SNS における
投稿件数推移分析のための
情報拡散モデル

平成 29 年度卒業論文

平成 30 年 2 月 1 日提出

千葉大学工学部都市環境システム学科

指導教員：塩田　茂雄

14T0265W

中島　圭佑

目次

第 1 章　序論 1
1.1　研究背景 1
1.2　研究目的 2
1.3　論文の構成 3
第 2 章　既存の数理モデル 4
2.1　SIS モデルと SIR モデル 4
2.2　既存研究 5
第 3 章　情報拡散モデル 6
3.1　モデルの説明 6
3.2　解析手法 7
3.3　相関項 10
3.3.1　独立近似 11
3.3.2　強相関近似 11
第 4 章　数値例 12
4.1　シミュレーション条件 12
4.2　シミュレーション(Facebook) 12
4.2.1　シミュレーション(Facebook:1 回) 12
4.2.2　シミュレーション(Facebook:1000 回) 15
4.3　シミュレーション(Twitter) 17
4.3.1　シミュレーション(Twitter:1 回) 17
4.3.2　シミュレーション(Twitter:1000 回) 20
4.4　解析評価 22
4.4.1　解析(Facebook) 22
4.4.2　解析(Twitter) 26
4.5　シミュレーションと解析の比較 29
4.5.1　シミュレーションと解析(Facebook) 29
4.5.2　シミュレーションと解析(Twitter) 33
第 5 章　結論 38
謝辞 39
参考文献 40

第 1 章　序論

1.1　研究背景

近年，Facebook や Twitter，Instagram や Google＋などのソーシャルネットワーキングサービス（Social Networking Service: SNS）が日常的なコミュニケーションツールとして普及し，人々の情報発信，収集の基盤となりつつある．

SNS の出現により，対面的なコミュニケーションが主流であった時代では想像できないような多様かつ膨大な情報が SNS を介して急速に拡散し，我々の生活に大きな影響を及ぼす時代となっている．その例を以下に示す．

・株価変動，予測

SNS 上の書き込みに影響されて株価が変動することがある．またその書き込みを分析して，株価変動を予測する試みが行われている[1].

・マーケティング

SNS を利用したバイラル(バズ)マーケティングも盛んに行われており，SNS 上で話題になること(トレンド入りすること)がヒットの条件となりつつある．

・ニュース

いわゆるマスメディアが報道しなかった事件，事故，不祥事などについて，SNS を通じて知るきっかけになり，情報拡散によって世論を動かすことができる(誰でも発信者になれる)．検閲がないことも重要である．

・人間関係

現実で属している集団(会社，学校，親戚など)だけでは出会えないような人々に，SNS を通じて出会うことができる．人間関係を広げ，新たな世界を開拓することにつながる．

このように，SNS は日常のあらゆる場面で活用され，社会に不可欠な存在となりつつある．そのため，SNS 上で得られる情報や SNS で観察される様々な現象を解析し，そこから有為な知見を得るための試みが多数なされている[1]，[3]〜[14].

1.2　研究目的

現実社会で人々が事象に関心を持つと，関連する書き込みが SNS に投稿され，単位時間あたりの投稿数がピークを迎えた後，書き込みは次第に減少して，やがて沈静化する現象（投稿件数のスパイク現象）が起きる．

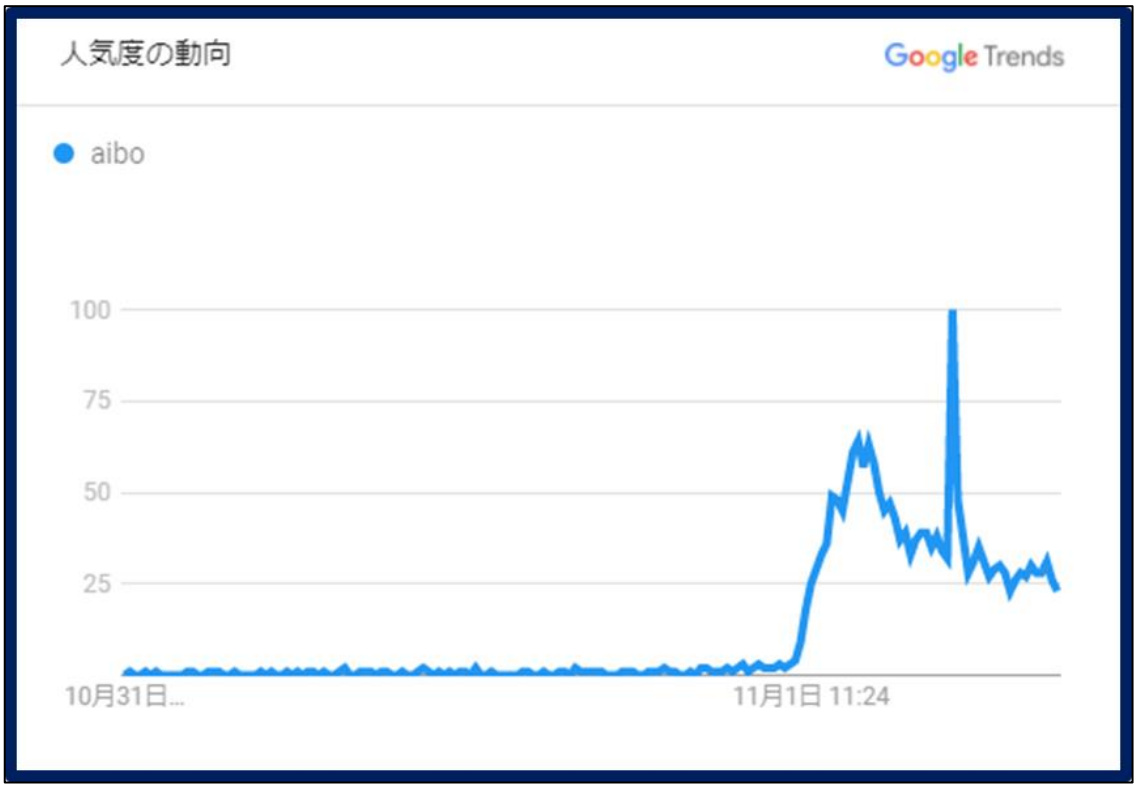

図 1　Google Trends[2]による「aibo」の検索回数の可視化

本研究では，これを現実社会で情報が人から人へ拡散して，最終的に広く知られるところとなり，それ以上の SNS への投稿が不要となるまでの過程，つまり現実社会における情報拡散過程が SNS 上に表出した現象と解釈する．

情報の拡散は感染症の流行と類似していることから，本研究では，この SNS への投稿件数の時間変化を，感染症の数理モデルとして使われている SIR モデル[3]により説明することを目的とする．

その際に，投稿件数の時間変化が，SNS のネットワーク構造（ネットワークトポロジー）にどのように依存するかを調べるため，隣接行列をパラメータとして含むモデルを採用する．さらに，モデルをシミュレーションで評価するだけでなく，SNS の投稿数の時間推移を解析的に評価する手法を検討する．

1.3　論文の構成

以下に，本論文の構成内容を述べる．

第 1 章　序論

本章であり，本研究の背景と目的，構成について述べた．

第 2 章　既存の数理モデル

既存の数理モデルの説明，及びそれらを利用した研究について述べる．

第 3 章　情報拡散モデル

本研究で用いた情報拡散モデルについて述べる．

第 4 章　数値例

SNS への投稿数の時間推移について，シミュレーション評価と解析評価を行い，モデルの妥当性を検討する．

第 5 章　結論

本研究の結論と今後の課題について述べる．

第 2 章　既存の数理モデル

2.1　SIS モデルと SIR モデル

感染症ウイルスや情報の拡散過程の分析に使われているモデルとして，SIS モデル，SIR モデルなどが挙げられる．SIS モデルとは，健康状態（Susceptible：状態 S）のノードが，感染状態（Infected：状態 I）の隣接ノードから確率的に感染して状態 I に遷移し(1)，また病気が自然治癒して状態 S に戻る(2)ということを繰り返すモデルである．

$$(1) S + I \rightarrow 2I$$
$$(2) I \rightarrow S$$

SIR モデルでは，病気が治癒すると免疫獲得状態（Recovered：状態 R）に遷移し，再び感染することはない．SIR モデルにおいて，感染と回復(隔離)の過程は，感染者 I との接触により，健康状態 S が感染者 I となる感染過程(3)と感染者 I が回復し(隔離され)，免疫保持者（隔離者）R となる回復(隔離)過程(4)の 2 つの過程に分離される．

$$(3) S + I \rightarrow 2I$$
$$(4) I \rightarrow R$$

接触率は健康状態 S と感染状態 I の両者の密度の積に比例し，感染率を β，回復(隔離)率を γ とすると，

$$\frac{dS(t)}{dt} = -\beta S(t) I(t)$$
$$\frac{dI(t)}{dt} = \beta S(t) I(t) - \gamma I(t)$$
$$\frac{dR(t)}{dt} = \gamma I(t)$$

となる．

式の和を取ると，

$$\frac{d}{dt}\big(S(t)+I(t)+R(t)\big)=0$$

であり，これは全体の人口$N(t)=S(t)+I(t)+R(t)$が一定値をとる保存則に対応している．この保存則により，本質的には 2 変数の方程式である．

このほかに，感染症の潜伏期間(Exposed)を考慮した SEIR モデルなどがある．

2.2 既存研究

SIS や SIR モデル等を用いた感染症の研究例として，ネットワークの構造を陽に考慮せずに，平均場近似を用いて現象論的なアプローチを試みるもの[4]，次数分布の情報を用いるもの[5]，隣接行列を用いてネットワークの構造を直接モデルの中に取り入れるもの[6]などがある．

この SIS モデルや SIR モデルを用いて，SNS 上の事象の分析を行った研究の例として[7]～[9]がある．Leskovec ら[7]は，被感染者をブログサイトとして，ブログサイト間の関連性が生成されるメカニズムを，SIS モデルを用いて分析している．Okada ら[8]は，本研究と同様，SNS の投稿数のスパイク（バースト的な書き込み）は情報拡散の帰結であるとして，ネットワーク構造を考慮しない巨視的な SIR モデルにより投稿数のスパイクの生成メカニズムを分析している．Cheng ら[9]は，SIR モデルに免疫獲得状態から感染状態へ遷移するメカニズムを取り入れることで，投稿数のスパイクが繰り返し生じる事象を説明できるとしている．

感染症や SNS 上の事象の分析に使われている SIS や SIR 以外の数理モデルの例として，IC（Independent Cascade）モデルや LT（Linear Threshold）モデルがある．IC モデルは，離散時間モデルを採用しており，ある時刻で情報を受け取ったノードiは，次の時刻で各隣接ノードに情報を転送する．ノードiの隣接ノードjは確率p_{ij}で情報を受け取ることができる（受け取りに失敗することがある）．

LT モデルも離散時間モデルを採用している．LT モデルでは，ネットワークを構成しているリンクにそれぞれ重みが付けられており，アクティブな情報を持つ隣接ノードからのリンクの重みが閾値を超えた場合，次の時点でそのノードはアクティブになる．IC モデルや LT モデルを用いた研究例として[10]～[13] がある．

その他の数理モデルを用いた研究例として，例えば，Matsubara ら[14]は，Blog の投稿数の多様なスパイクの形状を説明するために，ブログへの書き込みが更なる書き込みの誘発要因となること，また誘発の強さが書き込まれてからの時間とともに減衰すること等を特徴とするモデルを提案している．

第 3 章　情報拡散モデル

3.1　モデルの説明

本研究では，情報保持の状態を SIR に対応させたモデルを用いて，情報拡散過程を説明することを目的としている．本章では，その使用する情報拡散モデルと，その解析手法について説明する．

N 個のノードからなる有向ネットワークにおいて，1 つまたは複数のノードを起点として，情報が有向リンクを経由してネットワーク内の各ノードに拡散していく現象について考察する．ノードは以下のいずれかの状態を取る．

状態 0：情報を知らない（SIR モデルにおける Susceptible 状態）

状態 1：情報は知っているが，拡散しない
（SIR モデルにおける Infected 状態：感染力なし）

状態 2：情報を知っており，将来拡散予定
（SIR モデルにおける Infected 状態：感染力あり）

状態 3: 情報を知り，拡散済み（SIR モデルにおける Recovered 状態）

時刻 0 では情報発信元のノードは状態 2 にあり，それ以外の全てのノードは状態 0 にある．ノードiは隣接ノードから情報を受信することで，状態 0 から確率$1-q_i$で状態 1 に，確率q_iで状態 2 に遷移する(q_iを情報拡散確率と呼ぶ)．状態 0 から状態 1 に遷移した場合は，状態 1 に留まるが，別の隣接ノードから情報を受信すると確率q_iで状態 2 に遷移する．状態 2 に遷移後は，平均$1/\lambda_i$の指数分布に従う時間経過後に，全ての隣接ノードに同時に情報転送を行い，状態 3 に遷移する．状態 3 に遷移後は，そのまま状態 3 に留まる(図 2)．

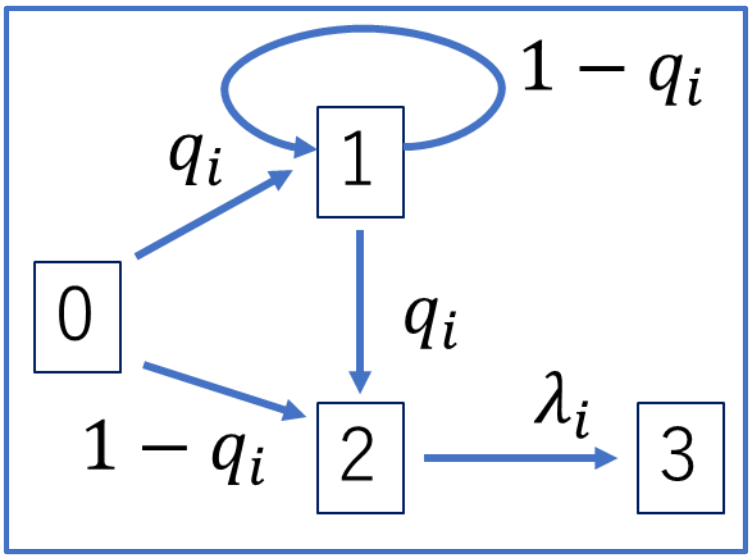

図 2　ノードiの状態遷移図

隣接行列を$A=\{a_{ij}\}$とし，ノードiからノードjへの情報転送によりノードjが（状態 0 から）状態 2 に遷移する場合に 1，遷移しない場合に 0 を取る確率変数をY_{ij}とする（$a_{ij}=0$のときは$Y_{ij}=0$とする）．時刻tにおいてノードiがとる状態の番号を$Z_i(t)$で表すこととすると，時刻tでのネットワークの状態は，確率変数$Y \stackrel{\text{def}}{=} \{Y_{ij}\}_{i,j=1,\ldots,N}$の値および各ノードの状態の組（$Z_1(t),Z_2(t),\ldots,Z_N(t)$）で記述され，（$Z_1(t),Z_2(t),\ldots,Z_N(t)$）の遷移は連続時間マルコフ連鎖に従う．

3.2　解析手法

本節では，シミュレーションに加えて，解析的にモデルを評価する手法について考察する．解析的な評価は，シミュレーションが定量評価であることに対し，定性的なモデルの理解に役立ち，また評価時間の短縮にも繋がる．

ネットワークの状態遷移を記述するマルコフ連鎖が既約で正再帰的であれば，いずれ全てのノードが状態 3 に遷移し，そこが定常状態となる．すなわち，本モデルの場合，定常状態は自明である．しかし，興味の対象は初期状態から定常状態への遷移の過程，つまり過渡特性にあり，（$Z_1(t),Z_2(t),\ldots,Z_N(t)$）の取り得る状態数は$4^N$（ノードごとに状態 0，1，2，3 の場合が存在する）であるので，（$Z_1(t)$，$Z_2(t),\ldots,Z_N(t)$）の遷移を厳密に追うことは，確率変数 Yの実現値が既知であっても困難である．

本節では，$p_i(t|Y) \stackrel{\text{def}}{=} E[X_i(t)|Y]$の時間変化に関する微分方程式を近似的に導出し，導出した微分方程式を数値的に解いて，ネットワークの状態遷移の過渡特性を解析的に評価することを試みる．

そのために，以下の確率変数を導入する．

$$X_i{}^{(k)}(t) = \begin{cases} 1 & Z_i{}^{(k)}(t) = k \\ 0 & otherwise \end{cases}$$

このとき，ノードiが時刻tで状態kにいる確率$p_i{}^{(k)}(t|Y)$は以下のように表される．

$$p_i{}^{(k)}(t|Y)(= \mathrm{E}\left[X_i{}^{(k)}(t)|Y\right])$$

ノードiの状態 1 への遷移は，ノードjがノードiと隣接し（ノードjからノードiへの有向リンクがあり），ノードjが状態 2 にあり，かつノードiが状態 0 にあるときに生じる．

ノードiの状態 2 への遷移は，ノードjがノードiと隣接し（ノードjからノードiへの有向リンクがあり），ノードjが状態 2 にあり，かつノードiが状態 0 または状態 1 にあるときに生じる．これを微分方程式で表すと，

$$\begin{aligned} \frac{dp_i{}^{(2)}(t|Y)}{dt} = &\sum_j a_{ji}\, Y_{ji}\lambda_j E\left[X_j{}^{(2)}(t)X_i{}^{(0)}(t)|Y\right] \left(0 \text{ から } 2 \text{ への遷移}\right) \\ &+ \sum_j a_{ji}\, Y_{ji}\lambda_j E\left[X_j{}^{(2)}(t)X_i{}^{(1)}(t)|Y\right] \left(1 \text{ から } 2 \text{ への遷移}\right) \\ &- \lambda_i E\left[X_i{}^{(2)}(t)|Y\right] \left(2 \text{ から } 3 \text{ への遷移}\right) \end{aligned}$$

また，ノードiの状態 2 から状態 3 への遷移は（隣接ノードの状態にかかわらず）遷移率λ_iで生じ，いったん状態 3 に遷移すると，以降，状態 3 に留まり続ける．

この微分方程式は

$$\frac{dp_i{}^{(3)}(t|Y)}{dt} = \lambda_i E\left[X_i{}^{(2)}(t)|Y\right] \left(2 \text{ から } 3 \text{ への遷移}\right)$$

時刻tでノードiが状態 2 か状態 3 にいる場合に 1，その他の場合に 0 を取る確率変数を$X_i(t)$で表すと，

$$X_i(t) \stackrel{\text{def}}{=} X_i{}^{(2)}(t) + X_i{}^{(3)}(t)$$
$$p_i(t) \stackrel{\text{def}}{=} p_i{}^{(2)}(t) + p_i{}^{(3)}(t)$$

また，

$$X_i^{(0)}(t) + X_i^{(1)}(t) + X_i^{(2)}(t) + X_i^{(3)}(t) = 1$$

このことから，

$$\begin{aligned}
&\frac{dp_i(t|Y)}{dt} \\
&= \frac{dp_i^{(2)}(t|Y)}{dt} + \frac{dp_i^{(3)}(t|Y)}{dt} \\
&= \sum_j a_{ji}\, Y_{ji}\lambda_j E\left[X_j^{(2)}(t)\left(X_i^{(0)}(t) + X_i^{(1)}(t)\right)|Y\right] \\
&= \sum_j a_{ji}\, Y_{ji}\lambda_j E\left[X_j^{(2)}(t)\left(1 - X_i(t)\right)|Y\right]
\end{aligned}$$

さらに，$a_{ji}Y_{ji}\lambda_j X_j^{(3)}(t)(1 - X_i(t))=0$ が成り立つ．

($Y_{ji} = 0$ のときは自明，$Y_{ji} = 1$ のときはノード i は状態 2 に遷移しているので，

$X_i(t) = 1$ となり式が成り立つ)

したがって，

$$\begin{aligned}
&\frac{dp_i(t|Y)}{dt} \\
&= \sum_j a_{ji}\, Y_{ji}\lambda_j E\left[X_j^{(2)}(t)\left(1 - X_i(t)\right)|Y\right] \\
&= \sum_j a_{ji}\, Y_{ji}\lambda_j E\left[(X_j^{(2)}(t) + X_j^{(3)}(t))(1 - X_i(t))|Y\right] \\
&= \sum_j a_{ji}\, Y_{ji}\lambda_j E\left[X_j(t)\left(1 - X_i(t)\right)|Y\right] \\
&= \sum_j a_{ji}\, Y_{ji}\lambda_j \left(p_j(t|Y) - E\left[X_j(t)X_i(t)|Y\right]\right) \quad \text{(A)}
\end{aligned}$$

が得られる．

本モデルにおける各ノードの情報転送は，SNS への書き込みの投稿に相当している．時刻 tまでに行われた情報転送の回数を$L(t)$とすると，

$$L(t) = \sum_i X_i^{(3)}(t).$$

であるから，

$$E[L(t)] = \sum_i E[X_i{}^{(3)}(t)] = \sum_i p_i{}^{(3)}(t).$$

ここで，ノードiが情報を受信した時刻をT_i，情報を受信してから情報転送までにかかった時間をτ_iとすると，

$$\begin{aligned} p_i{}^{(3)}(t|Y) &= P(T_i + \tau_i \le t|Y) \\ &= \int_0^t P(T_i \le t - s|Y) P(s \le \tau_i < s + ds) \\ &= \int_0^t \lambda_i P(T_i \le t - s|Y) e^{-\lambda_i s}\, ds \\ &= \int_0^t \lambda_i p_i(t - s|Y) e^{-\lambda_i s}\, ds = \int_0^t \lambda_i p_i(s|Y) e^{-\lambda_i (t-s)}\, ds \end{aligned}$$

したがって，

$$E[L(t)] = \textstyle\sum_i \int_0^t p_i(s|Y)\, e^{-\lambda_i (t-s)} ds. \quad \text{(B)}$$

すなわち，(A)を数値的に解いて$p_i(t|Y)$の値を求めることができれば，(B)から投稿数の期待値の時間変化を計算することができる．

(A)には$p_j(t|Y)$(ノードiの隣接ノード)の項が含まれているため，$E[L(t)]$を求めるためには，全ノードの情報保持確率$p_i(t|Y)$について連立微分方程式を立てて解いていく必要がある．そのためには，$E\left[X_j(t)X_i(t)|Y\right]$を明らかにする必要があるが，これは厳密に解くことはできないので, 何らかの仮定を用いる必要がある(この項の扱いについては3.3節で述べる).

解析評価を行うためには，$E[X_i(t)|Y] = p_i(t|Y)$を求める必要があるが，全ての確率変数 Y の実現値について$p_i(t|Y)$を求めて平均を取るのは困難である(1 つのリンクにつきY_{ij}は 0 か 1 の値を取るので， Yは 2 のリンク数乗存在する)ので，本研究ではランダムに Yの実現値を 1 つ選んで$p_i(t|Y)$を計算し，その結果を$E[X_i(t)|Y]$の近似値として用いることを考える．

3.3 相関項

微分方程式(A)の右辺に現れる$E\left[X_j(t)X_i(t)|Y\right]$（以下，相関項）は既知ではないので，(A)から$p_i(t|Y)$を数値的に計算するためには，何らかの仮定によって$E\left[X_j(t)X_i(t)|Y\right]$を$p_i(t|Y) = E[X_i(t)|Y]$と$p_j(t|Y) = E\left[X_j(t)|Y\right]$の関数で表現する必要がある．本研究では 2 種類の近似を用いて，(A)から$p_i(t|Y)$を数値的に計算する方法を検討する．

3.3.1　独立近似

(A)において，$E[X_j(t)X_i(t)|Y] = p_j(t|Y)p_i(t|Y)$ とする近似を「独立近似」と呼ぶ．独立近似は巨視的（現象論的）に情報拡散過程を解析する研究でしばしば用いられる仮定であり，これは情報保持確率の上限を与える．この近似の上では，情報拡散は最も速く進行する．独立近似のもとでは，情報保持確率の大小にかかわらず，任意の方向で情報の転送が生じる．

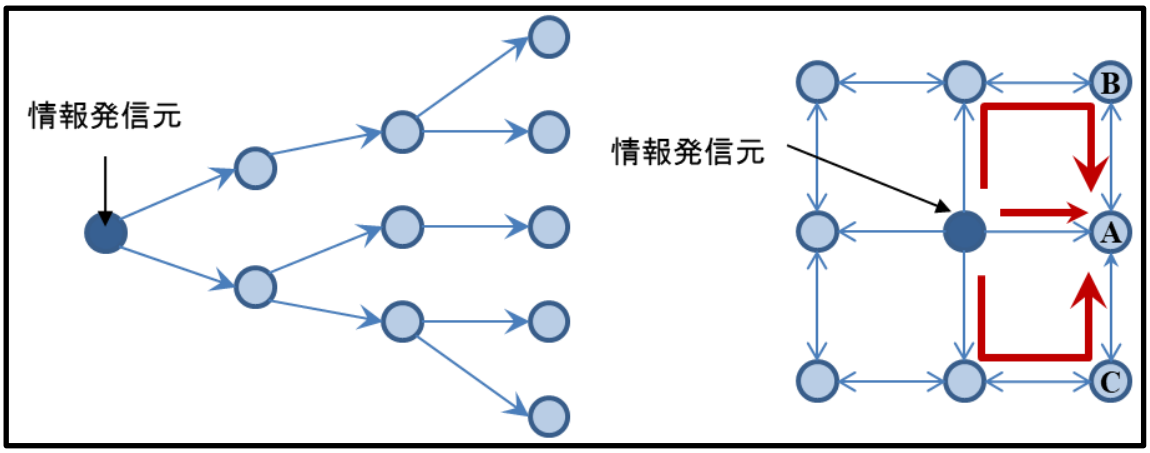

図３　ネットワークトポロジー例

このため独立近似のもとでは，例えば図３の左側のツリー構造のネットワークにおいて，情報の逆流（右側から左側への情報の流れ）が生じる．また，図３の右側の格子状のネットワークにおいては，ノードＢやノードＣを経由して，ノードＡに情報が到達する可能性を実際よりも過大に評価する傾向がある．

3.3.2　強相関近似

一方，(A)において $E[X_j(t)X_i(t)|Y] = min\{p_j(t|Y),\ p_i(t|Y)\}$ とする近似を「強相関近似」と呼ぶ．これは情報保持確率の下限を与える．この近似の上では，情報拡散は最も遅く進行する．強相関近似のもとでは，情報は情報保持確率の高いノードから情報保持確率の低いノードにしか流れない．

図 3 の左側のツリー構造のネットワークの場合，情報は情報発信元から下流のノードに一方向に流れる．このような場合は強相関近似が厳密に成り立つ．一方，右側の格子状のネットワークの場合，一般にノードＡの情報保持確率はノードＢやノードＣの情報保持確率より高いと考えられるが，情報発信元からノードＢやノードＣを経由して，ノードＡに情報が到達する可能性があるので，強相関近似は厳密には成り立たない．

第 4 章　数値例

本章では，情報拡散過程において，第 3 章で説明した情報拡散モデルを用いてシミュレーションと解析を行った結果を示す．

4.1　シミュレーション条件

本節では，シミュレーションと解析をどのような条件で行ったかを述べる．

本研究では，インターネット上に公開されている Facebook，Twitter のネットワークトポロジーデータ[15]を使用して，情報拡散のシミュレーション実験を行った．

Facebook のトポロジーデータはノードが 4039 個，リンクが 88234 本(無向グラフ)，Twitter のトポロジーデータはノードが 10000 個，リンクが 138678 本(有向グラフ)のものをそれぞれ使用した．

λ_i＝1として，1 つのノードから情報を発信し，投稿数（情報拡散回数）の時間推移を評価した．

4.2　シミュレーション(Facebook)

Facebook のトポロジーデータ上で，発信ノード次数，情報拡散確率を変えてシミュレーションを行った．

4.2.1　シミュレーション(Facebook:1 回)

情報拡散シミュレーションを 1 回ずつ行い，その結果を 3 パターンずつ示す．発信ノード次数は 30 として，情報拡散確率を変えてシミュレーションを行った．

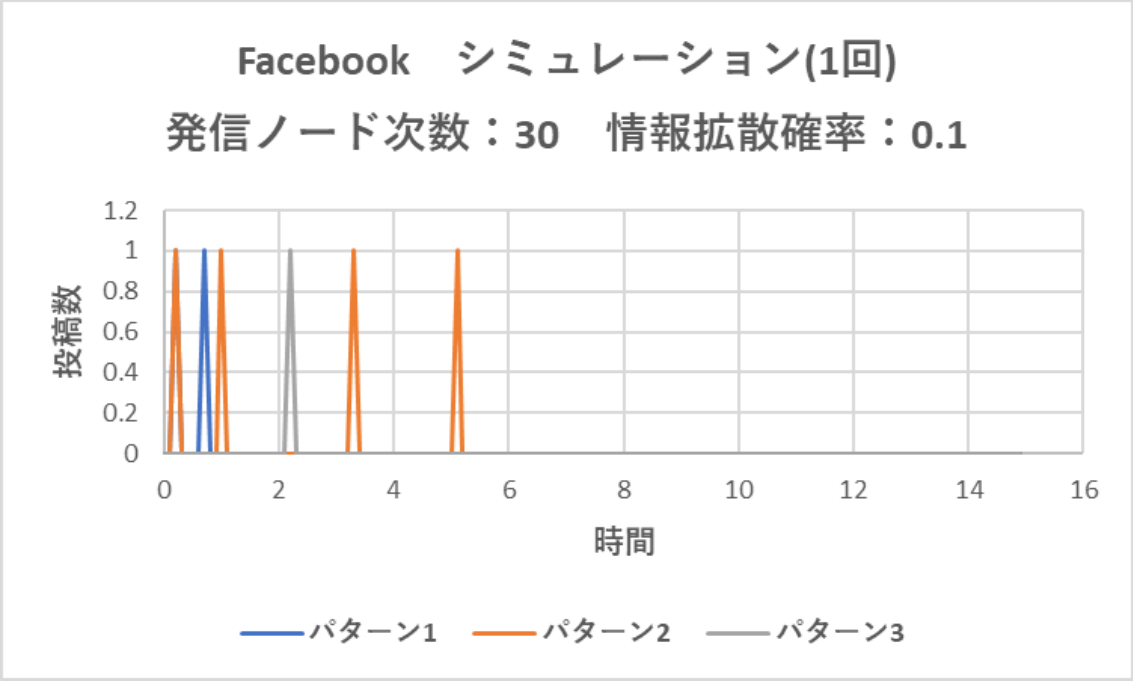

図 4　投稿数の時間推移シミュレーション 1 回(Facebook：情報拡散確率 0.1)

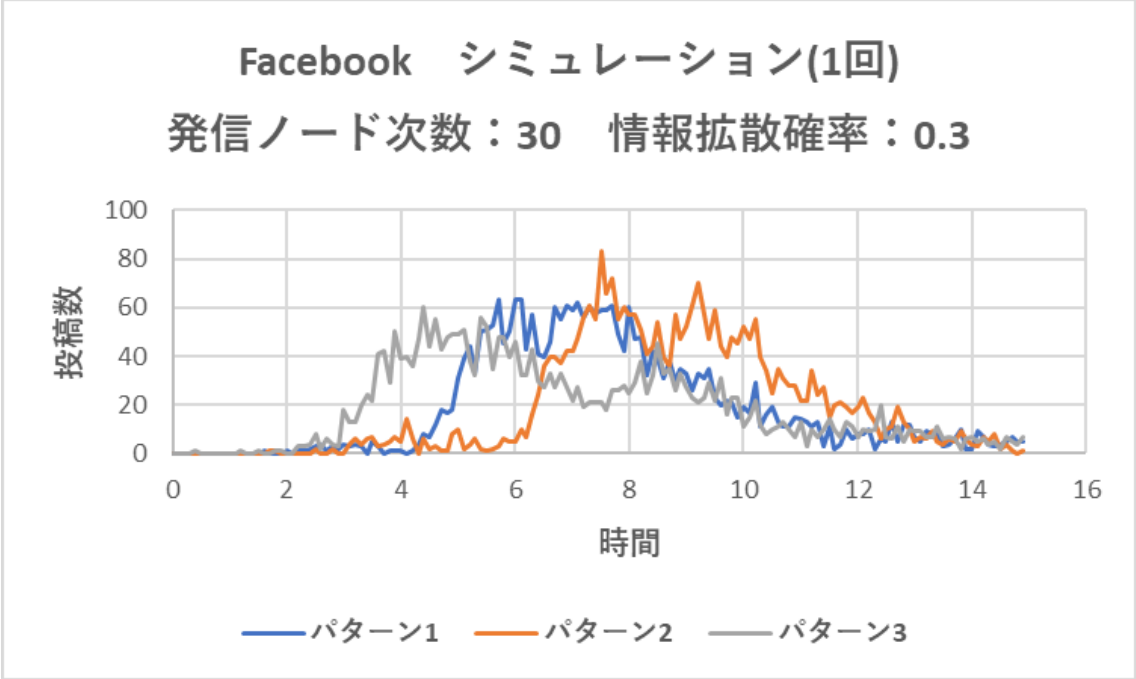

図 5　投稿数の時間推移シミュレーション 1 回(Facebook：情報拡散確率 0.3)

図 6　投稿数の時間推移シミュレーション 1 回(Facebook：情報拡散確率 0.5)

図 7　投稿数の時間推移シミュレーション 1 回(Facebook：情報拡散確率 1)

図 4～図 7 の結果から，シミュレーション 1 回ごとに情報拡散の様子が大きく違うことが分かる．特に，情報拡散確率 0.1 のときは顕著である．本研究では，シミュレーションを 1000 回行ったものを平均し，解析結果と比較する．

4.2.2　シミュレーション(Facebook:1000 回)

各発信ノードに対して，情報拡散確率(ノードが状態 2 に遷移する確率)がそれぞれ 0.1, 0.3，0.5，1.0 の時の結果を示した.

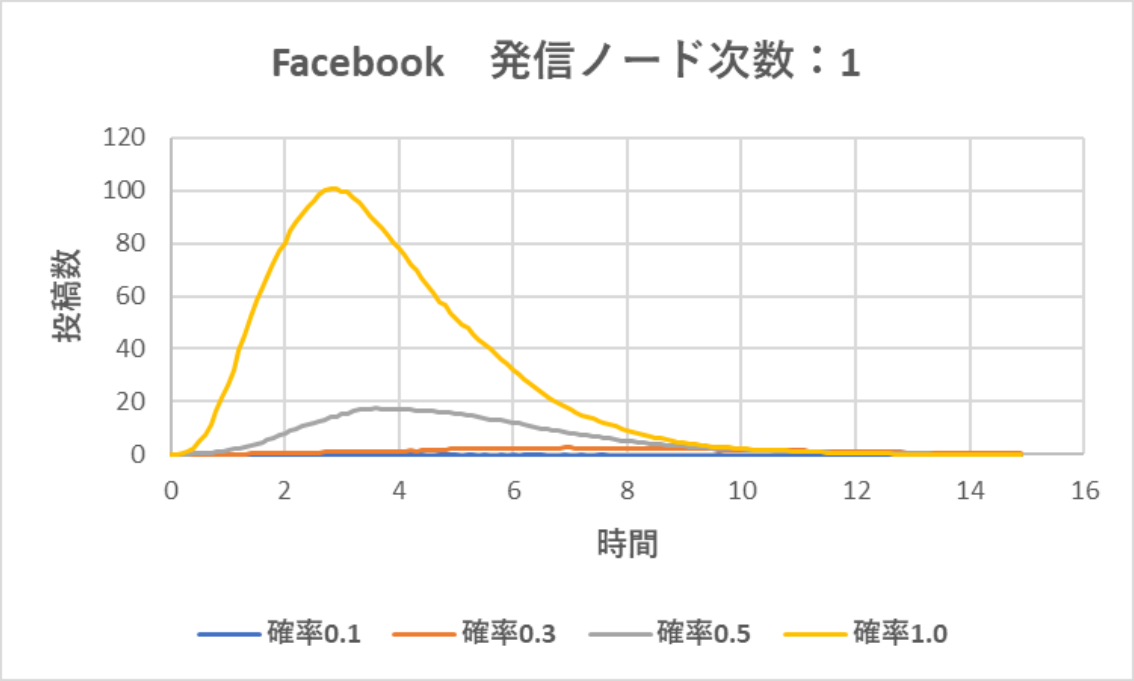

図 8　投稿数の時間推移シミュレーション(Facebook：発信ノード次数＝1)

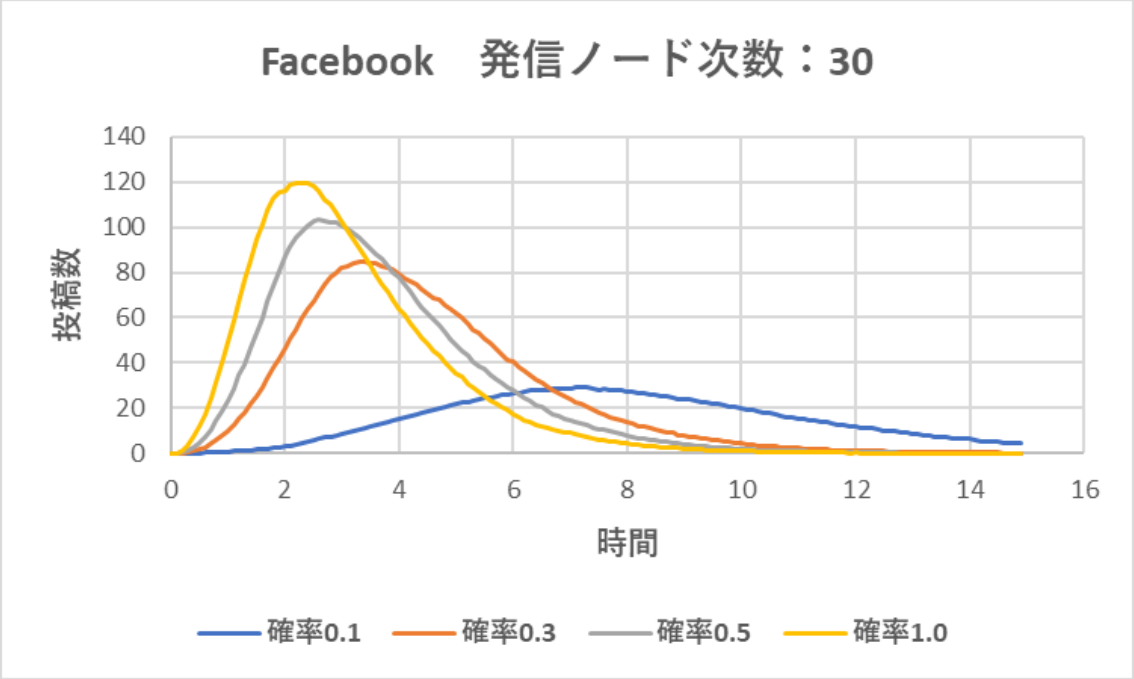

図 9　投稿数の時間推移シミュレーション(Facebook：発信ノード次数＝30)

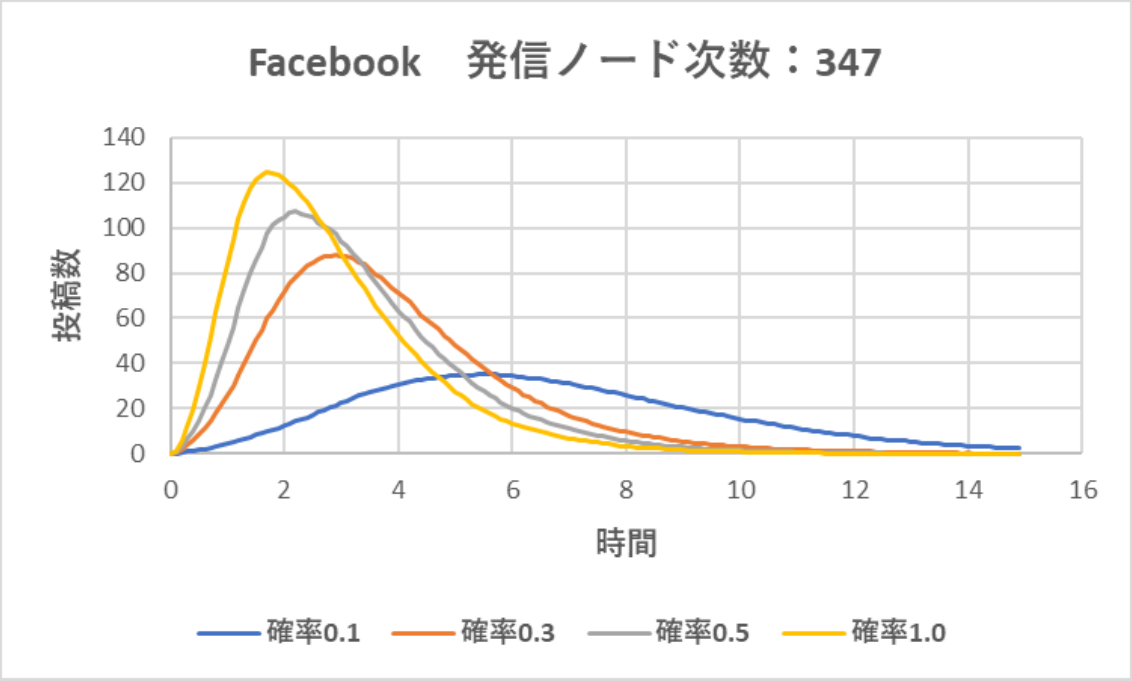

図 10　投稿数の時間推移シミュレーション(Facebook：発信ノード次数＝347)

情報拡散確率が低下するほど投稿数の総数が少なくなり，情報拡散のピーク(単位時間当たりの投稿数が最も大きくなる部分)が後に来ることが確認できる．また，発信ノード次数が大きくなるほど，情報拡散のピークが先に来て，スパイクが急峻になることが確認できる．

4.3　シミュレーション(Twitter)

Twitter のトポロジーデータを用いて，同様のシミュレーションを行った結果を示す．Twitter のネットワークトポロジーは有向グラフであるため，ノードの出次数と入次数に留意する必要がある．

4.3.1　シミュレーション(Twitter:1 回)

Facebook と同様に，情報拡散シミュレーションを 1 回ずつ行い，その結果を 3 パターンずつ示す．発信ノード次数は 41 として，情報拡散確率を変えてシミュレーションを行った．

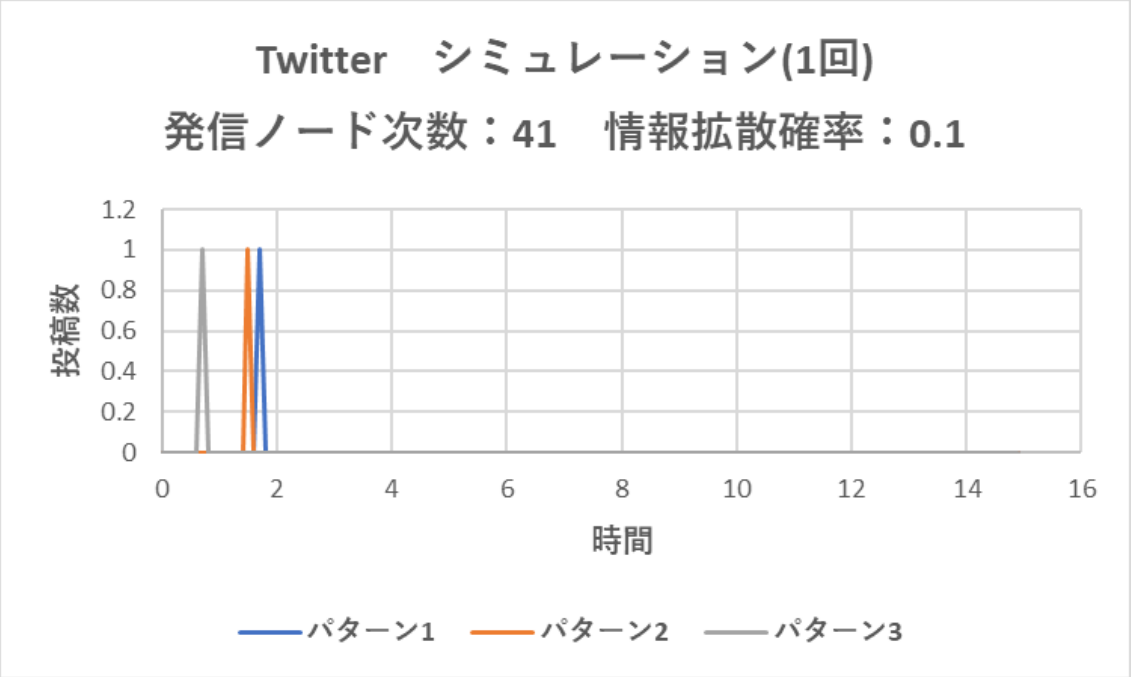

図 11　投稿数の時間推移シミュレーション 1 回(Twitter：情報拡散確率 0.1)

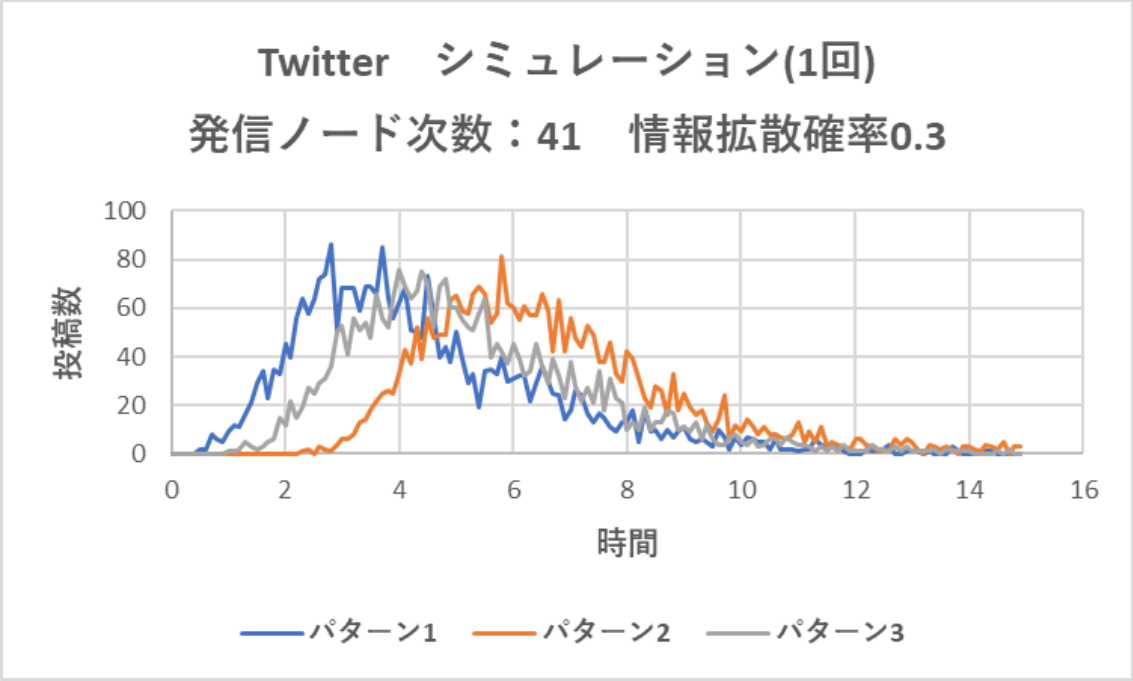

図 12　投稿数の時間推移シミュレーション 1 回(Twitter：情報拡散確率 0.3)

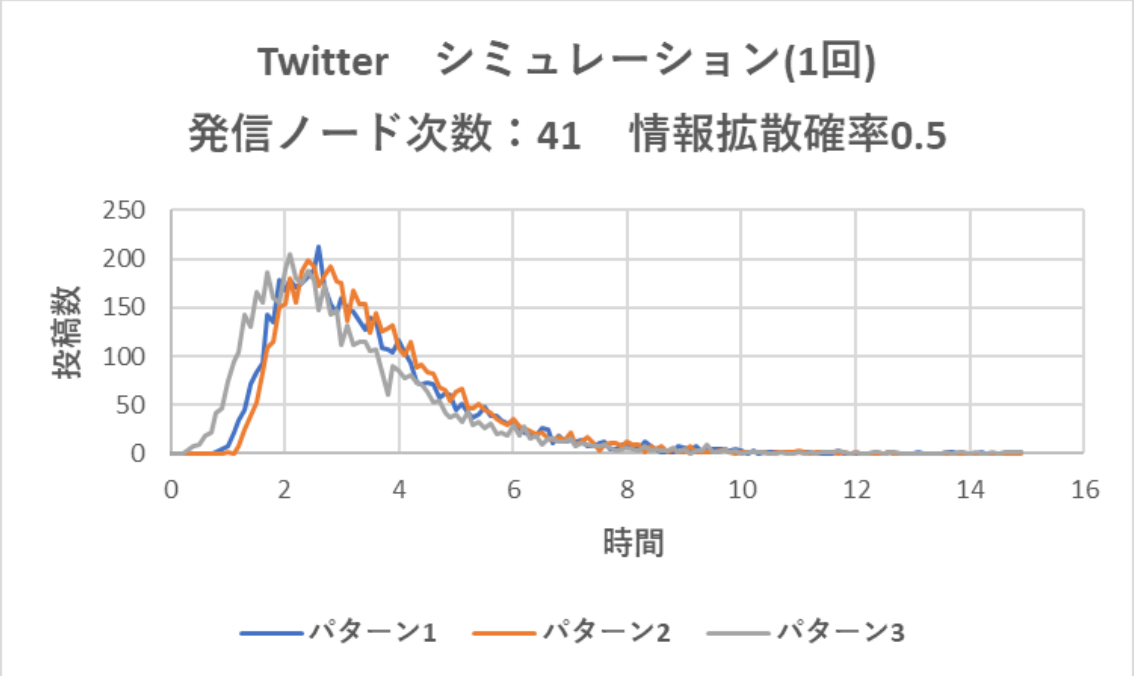

図 13　投稿数の時間推移シミュレーション 1 回(Twitter：情報拡散確率 0.5)

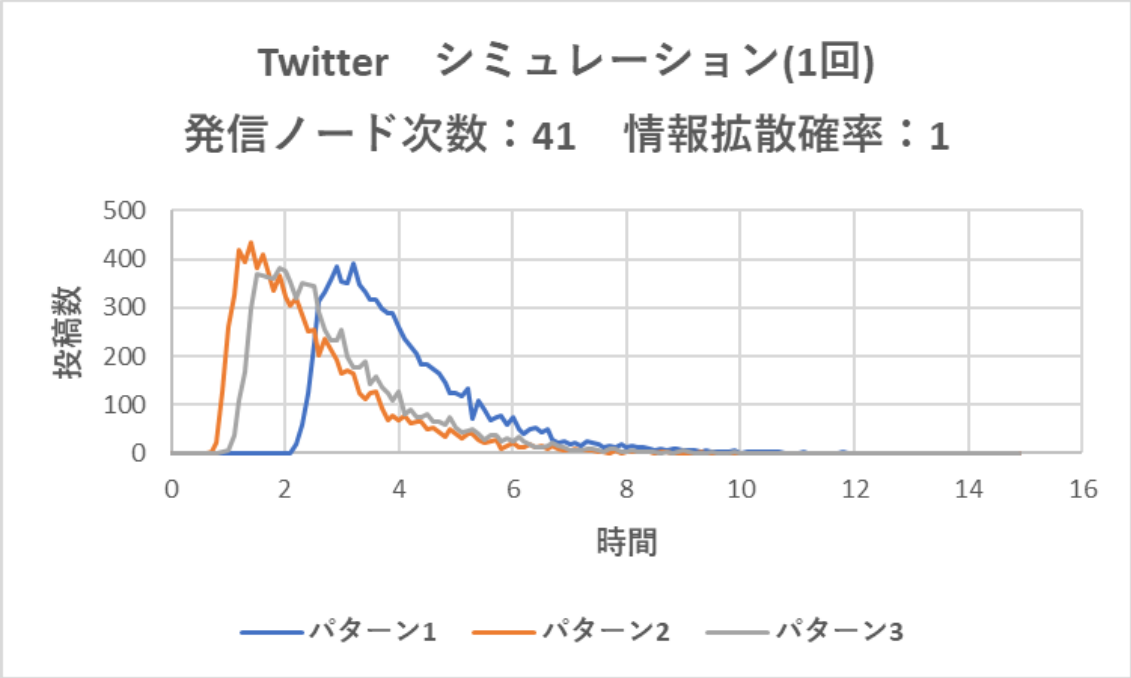

図 14　投稿数の時間推移シミュレーション 1 回(Twitter：情報拡散確率 1)

図 11～図 14 の結果から，Facebook のときと同様にシミュレーション 1 回ごとに情報拡散の様子が大きく違うことが分かる．　Facebook のときと同様，本研究ではシミュレーションを 1000 回行ったものを平均し，解析結果と比較する．

4.3.2　シミュレーション(Twitter:1000 回)

各発信ノードに対して，情報拡散確率(ノードが状態 2 に遷移する確率)がそれぞれ 0.1，0.3，0.5，1.0 の時の結果を示した．

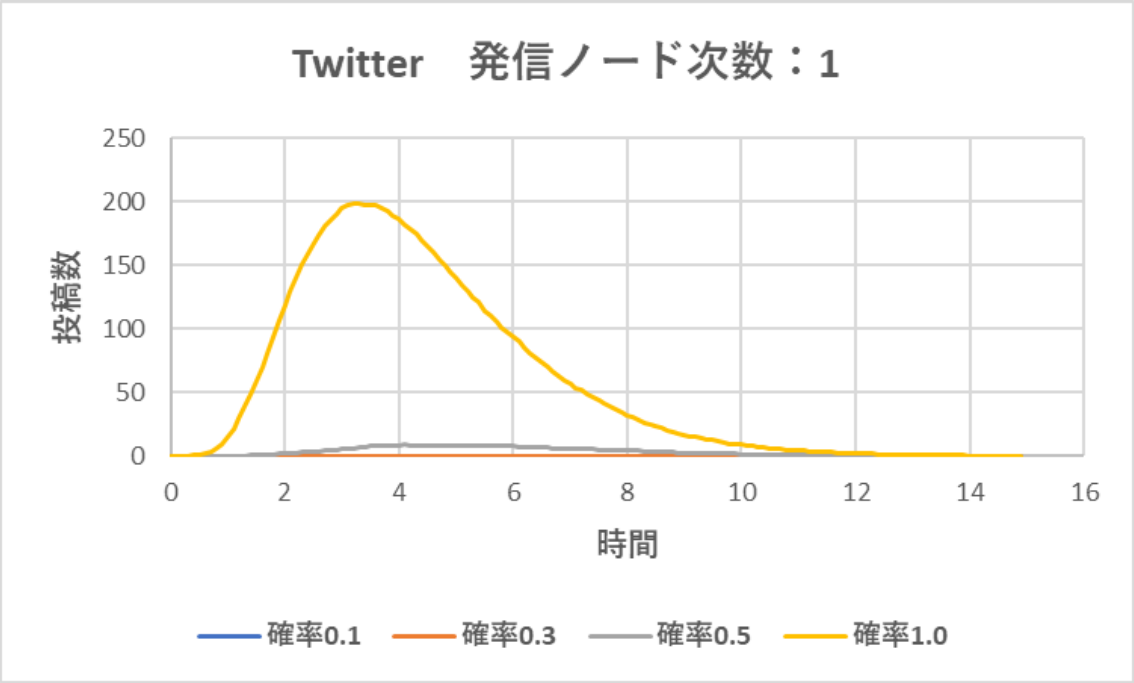

図 15　投稿数の時間推移シミュレーション(Twitter：発信ノード出次数＝1)

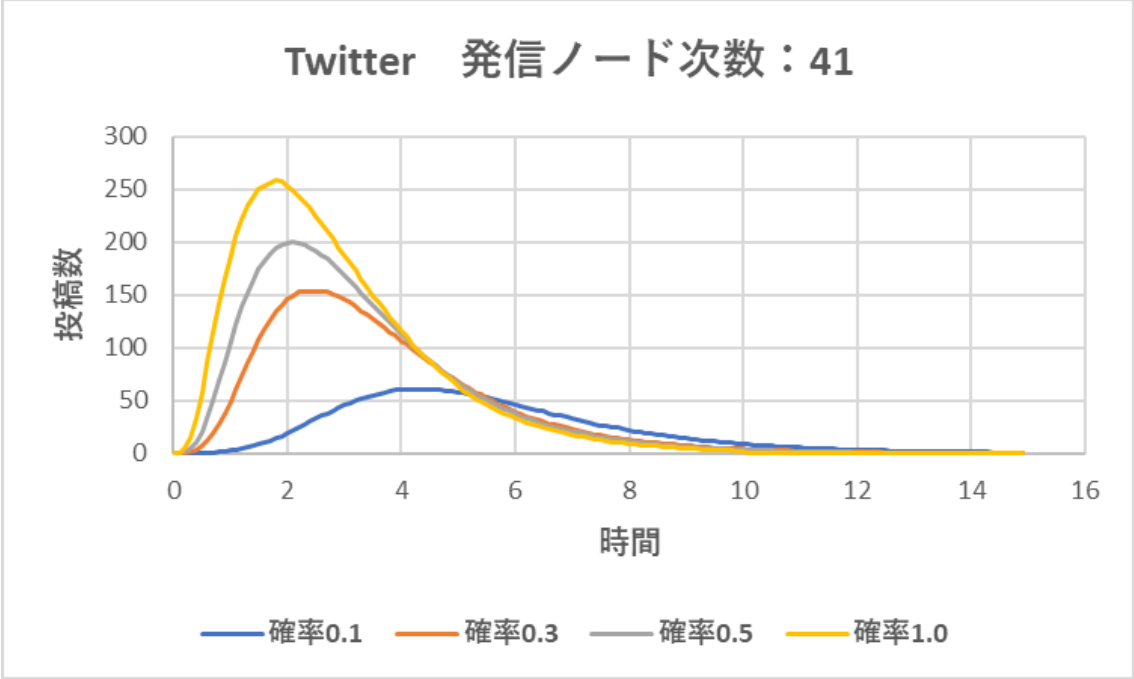

図 16　投稿数の時間推移シミュレーション(Twitter：発信ノード出次数＝41)

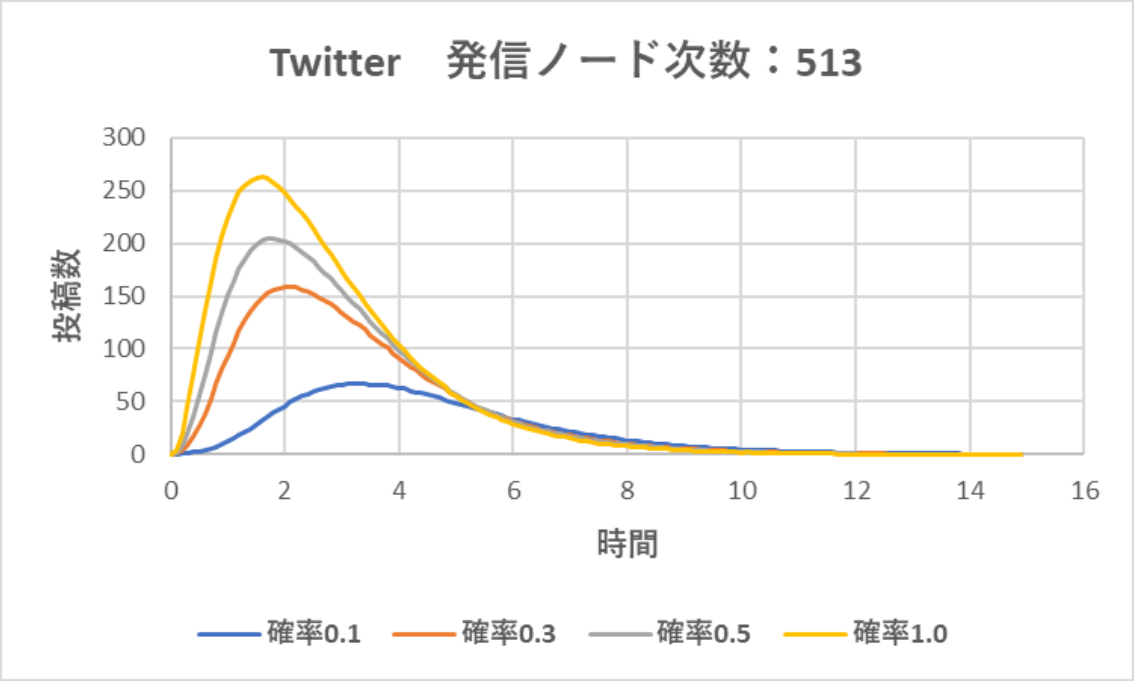

図 17　投稿数の時間推移シミュレーション(Facebook：発信ノード出次数＝513)

Facebook の投稿数シミュレーションと同様に，情報拡散確率の低下に伴い総投稿数が減少しピークが後に来ることが分かる．また，発信ノード次数が増加すると，ピークが先に来てスパイクが急峻になる点も同一である．しかし，Facebook の結果と比較すると，そのスパイクへの影響は小さい(発信ノード次数：41 と発信ノード次数：513 の結果において，ピークの時間は 2 から大きく変化せず，また情報拡散確率を変えてもピークは大きく変化しなかった)．

このことから，Twitter のユーザーごとに次数分布のばらつきがあることが推測できる．つまり，影響力の大きい人があらかた情報を拡散してしまい，そこで情報拡散の大部分は終わってしまう．Twitter は有向グラフ(フォロー：フォロワーが 1：1 ではない)なので，情報拡散をする際にはいかに影響力のある人(インフルエンサー)まで情報を持っていくかが重要になると考えられる．

4.4　解析評価

3.2 節で説明した解析手法を用いて，投稿数の時間推移について解析評価を行った．本節では 3.2 節で述べた「ランダムに Yの実現値を 1 つ選んで計算すること」の合理性を検証するため，Yをそれぞれランダムに 3 通り選んで計算を行い，その差異を比較した．

4.4.1　解析(Facebook)

Facebook のトポロジーデータ上で，発信ノード次数 30 として投稿数の解析評価を行った．情報拡散確率は 0.1，0.3，0.5 として，独立近似，強相関近似をそれぞれ適用した結果を示した．

図 18　投稿数の時間推移解析：独立近似(Facebook：情報拡散確率 0.1)

図 19　投稿数の時間推移解析：強相関近似(Facebook：情報拡散確率 0.1)

図 20　投稿数の時間推移解析：独立近似(Facebook：情報拡散確率 0.3)

図 21　投稿数の時間推移解析：強相関近似(Facebook：情報拡散確率 0.3)

図 22　投稿数の時間推移解析：独立近似(Facebook：情報拡散確率 0.5)

図 23　投稿数の時間推移解析：強相関近似(Facebook：情報拡散確率 0.5)

図 10, 図 11 の結果から，情報拡散確率 0.1 のときは Yの選び方によって解析結果に大きく差が出てしまい，ランダムに Y を選んで計算することは合理的ではないと考えられる．逆に，情報拡散確率 0.3，0.5 のときは解析結果に大きな差が見られず，Y をランダムに選んで計算することは合理的であると考えられる．

情報拡散確率 0.1 のときに解析結果に大きな差が出るのは，多くのリンクが無効になり(a_{ij}= 0 のときはY_{ij}= 0 とすることから，情報拡散確率 0.1 のときは 90%のリンクが無効になる)，グラフそのものが非連結グラフになってしまうからだと考えられる．そのため，ほとんど情報拡散が起こらず，投稿数が増えなかったと考えられる．

この「情報拡散確率が最低どれくらいあれば情報が拡散していくのか」という点については，今後の研究で検討する必要があるといえる．

4.4.2　解析(Twitter)

同様に Twitter のトポロジーデータ上で投稿数の解析評価を行った．発信ノード次数は 41 とした．

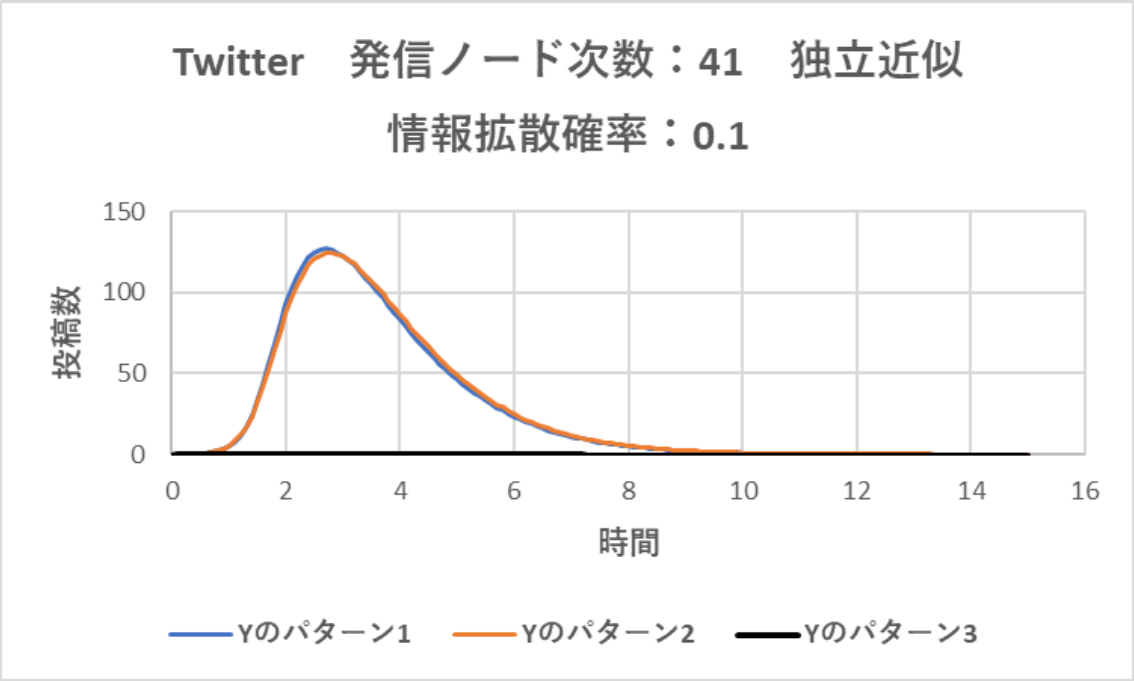

図 24　投稿数の時間推移解析：独立近似(Twitter：情報拡散確率 0.1)

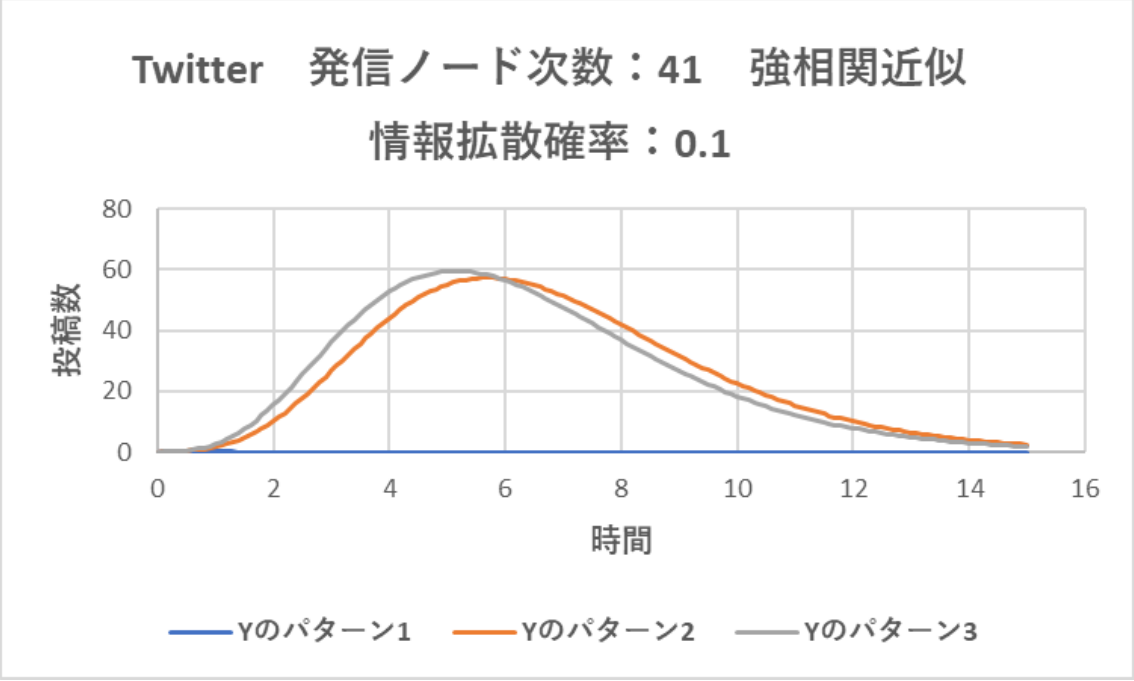

図 25　投稿数の時間推移解析：強相関近似(Twitter：情報拡散確率 0.1)

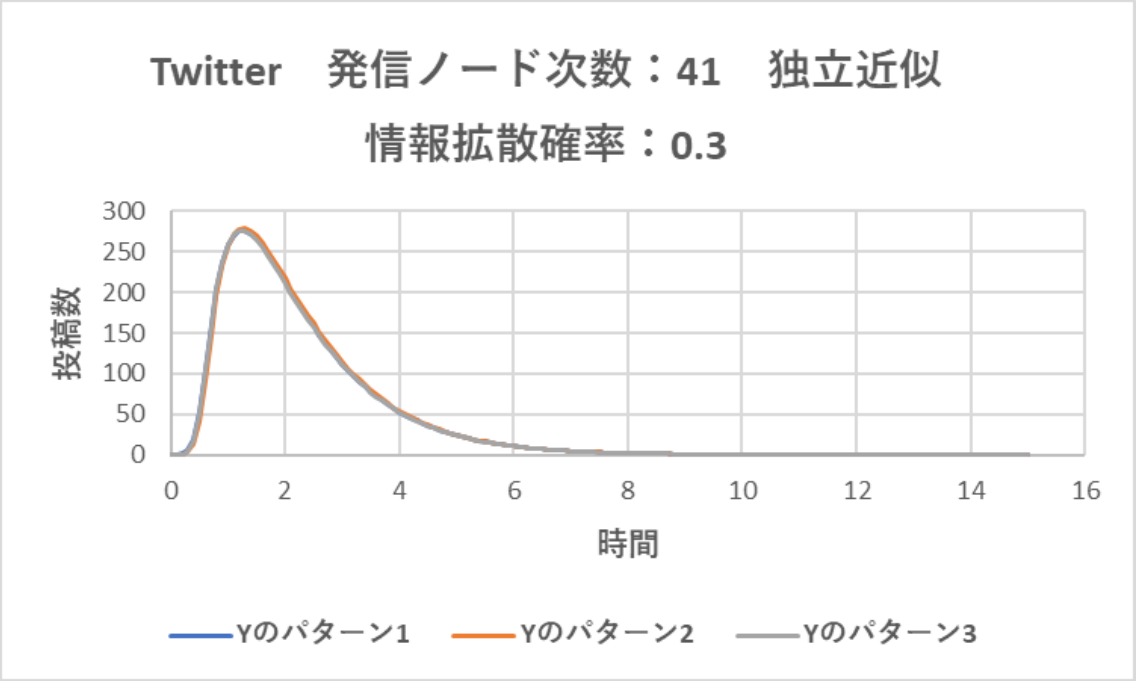

図 26　投稿数の時間推移解析：独立近似(Twitter：情報拡散確率 0.3)

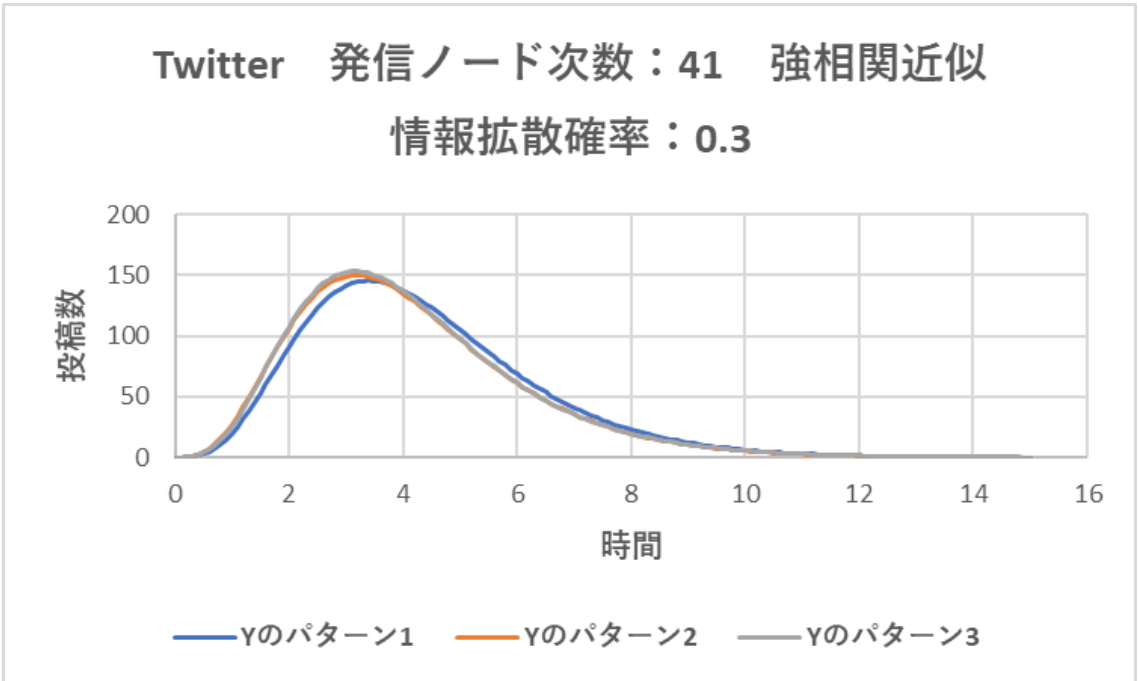

図 27　投稿数の時間推移解析：強相関近似(Twitter：情報拡散確率 0.3)

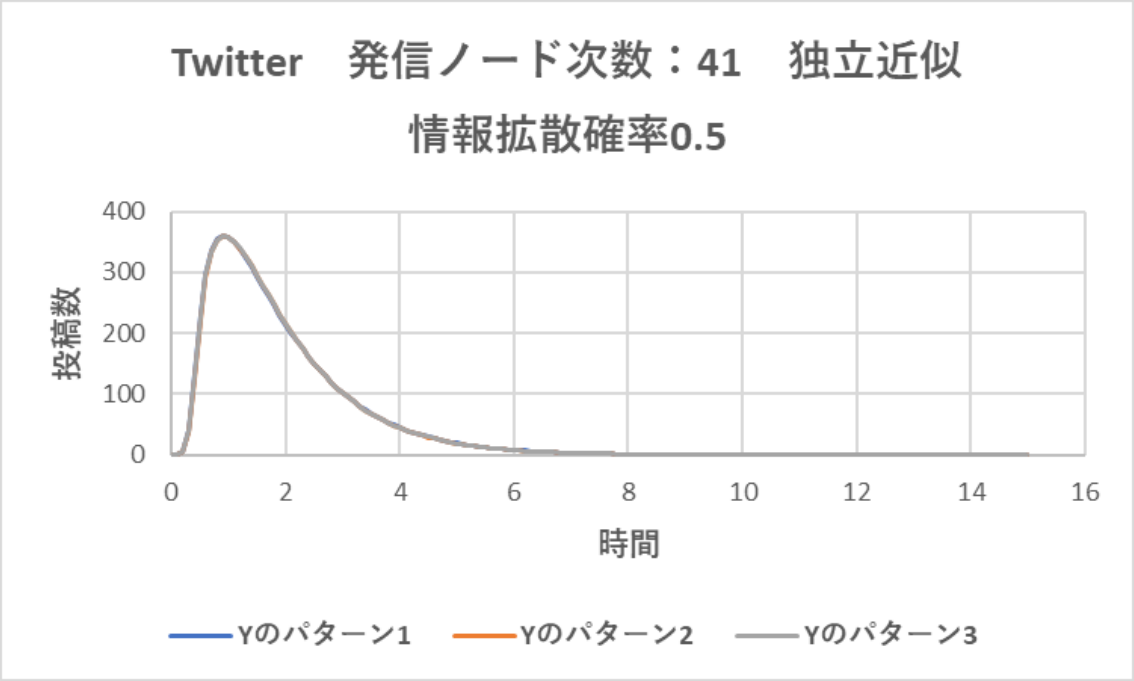

図 28　投稿数の時間推移解析：独立近似(Twitter：情報拡散確率 0.5)

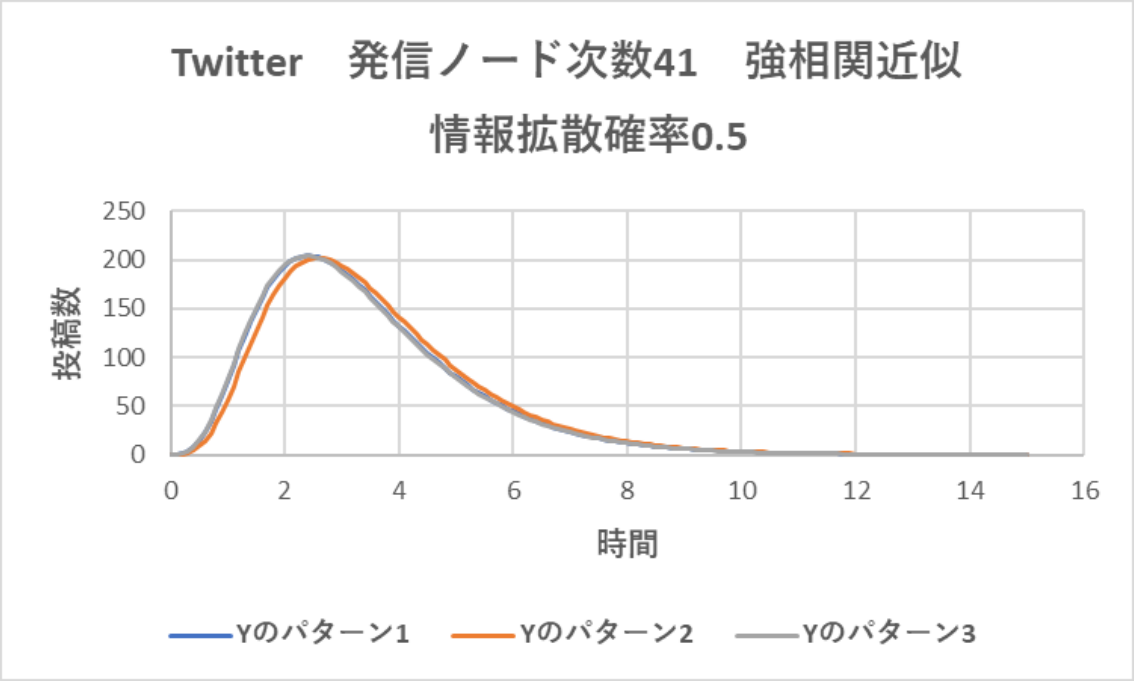

図 29　投稿数の時間推移解析：強相関近似(Twitter：情報拡散確率 0.5)

Facebook の解析結果と同様に，情報拡散確率 0.1 の場合，Y の選び方によって解析結果に大きな差が出る．情報拡散確率 0.3，0.5 のときは 3 パターンの Y でほぼ同様の結果が出るので，解析に利用できると考えられる．

4.5　シミュレーションと解析の比較

本節では，4.3 節，4.4 節で示したシミュレーション結果と解析結果を比較し，近似を用いて解析を行うことの妥当性を検証する．4.4 節で示した通り，情報拡散確率 0.1 の場合は Y の選び方によって結果に大きな差が出るため，解析結果を参考にすることはできない．

そのため，本節では情報拡散確率 0.3，0.5，1.0 の場合において，シミュレーション結果と解析結果を比較していく．

4.5.1　シミュレーションと解析(Facebook)

Facebook において，発信ノード次数：30 と発信ノード次数：1045 の結果をそれぞれ情報拡散確率 0.3，0.5，1.0 の場合ごとに示す．

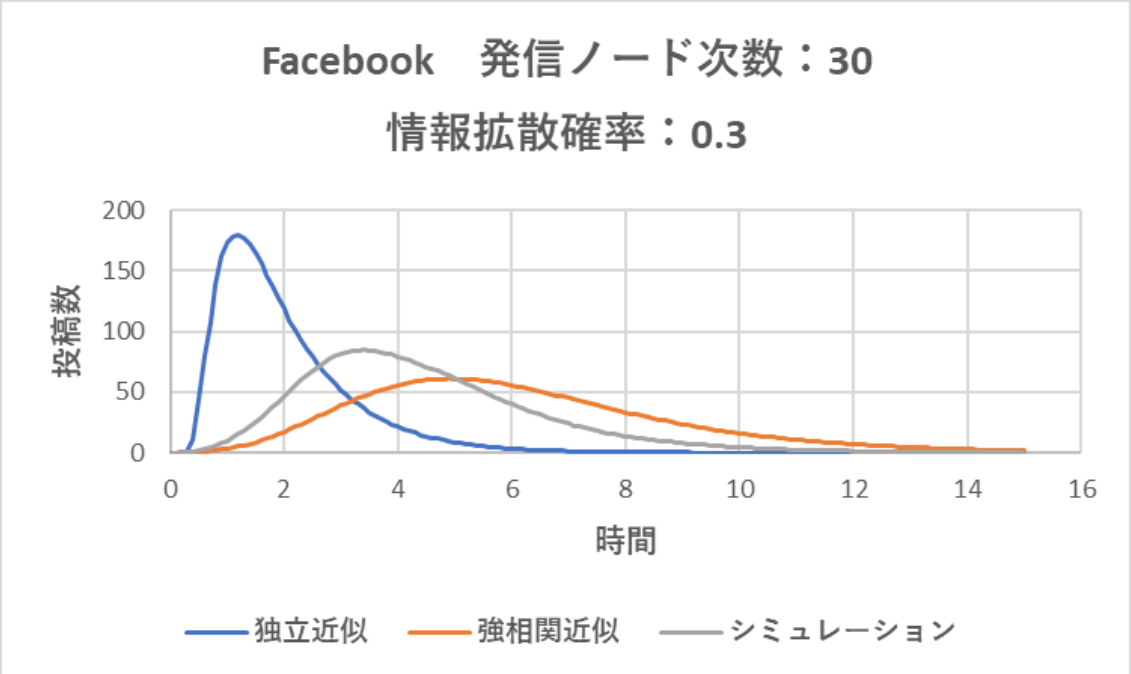

図 30　投稿数の時間推移（Facebook：発信ノード次数：30　情報拡散確率 0.3）

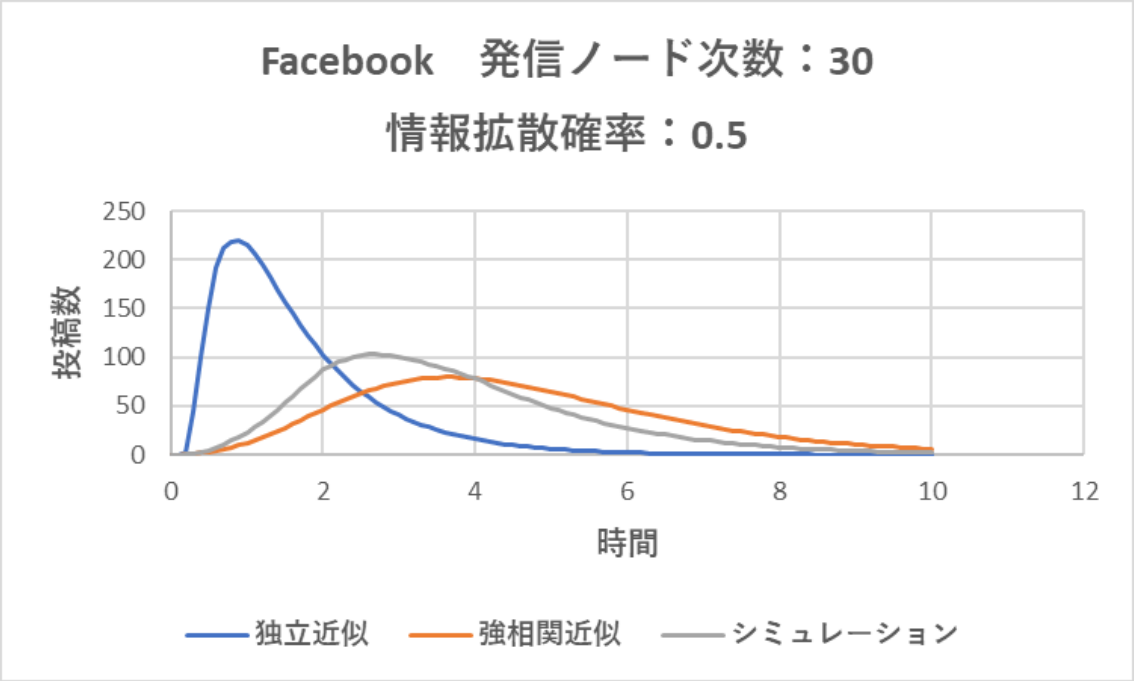

図 31　投稿数の時間推移（Facebook：発信ノード次数：30　情報拡散確率 0.5）

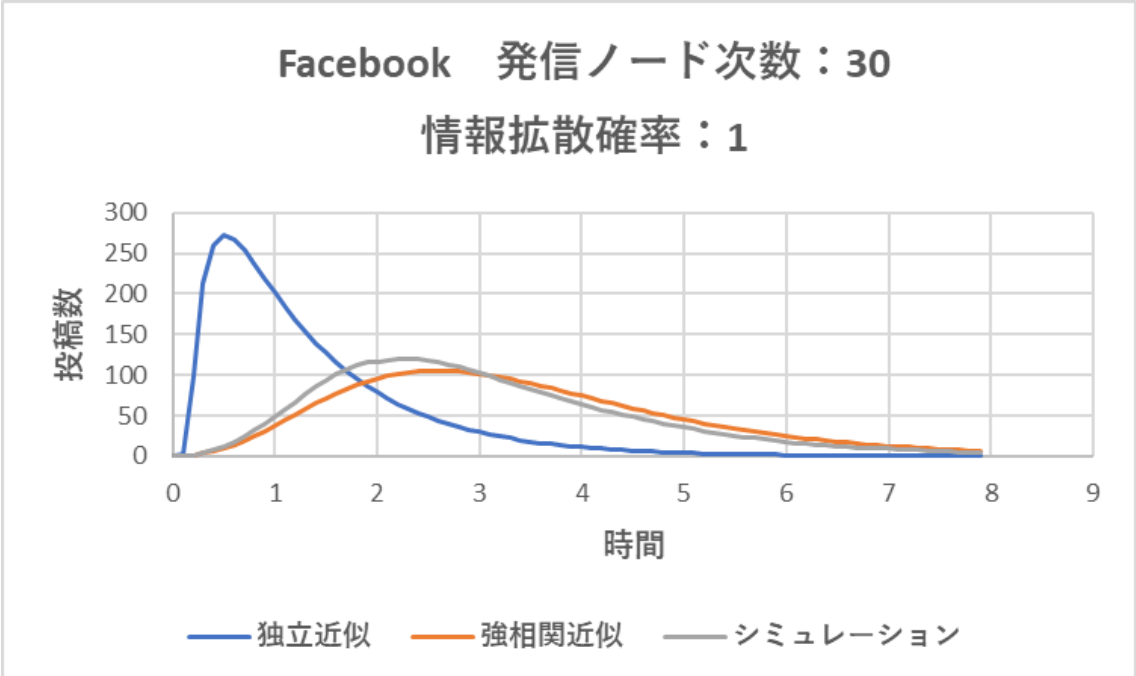

図 32　投稿数の時間推移（Facebook：発信ノード次数：30　情報拡散確率 1）

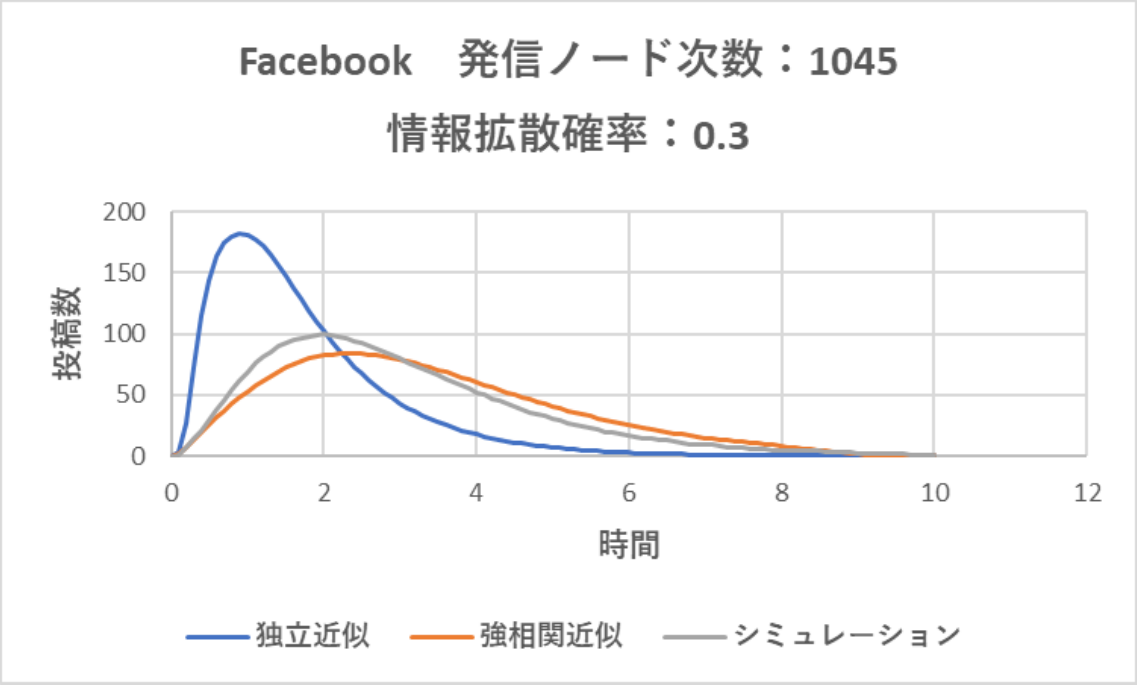

図 33　投稿数の時間推移（Facebook：発信ノード次数：1045　情報拡散確率 0.3）

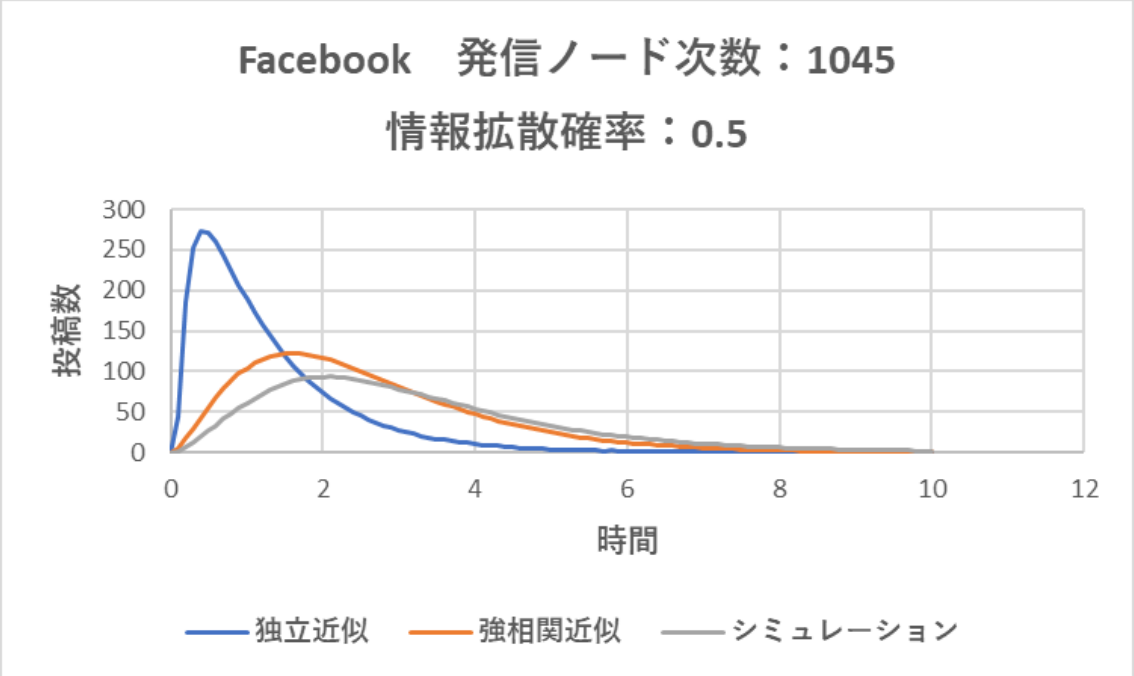

図 34　投稿数の時間推移（Facebook：発信ノード次数：1045　情報拡散確率 0.5）

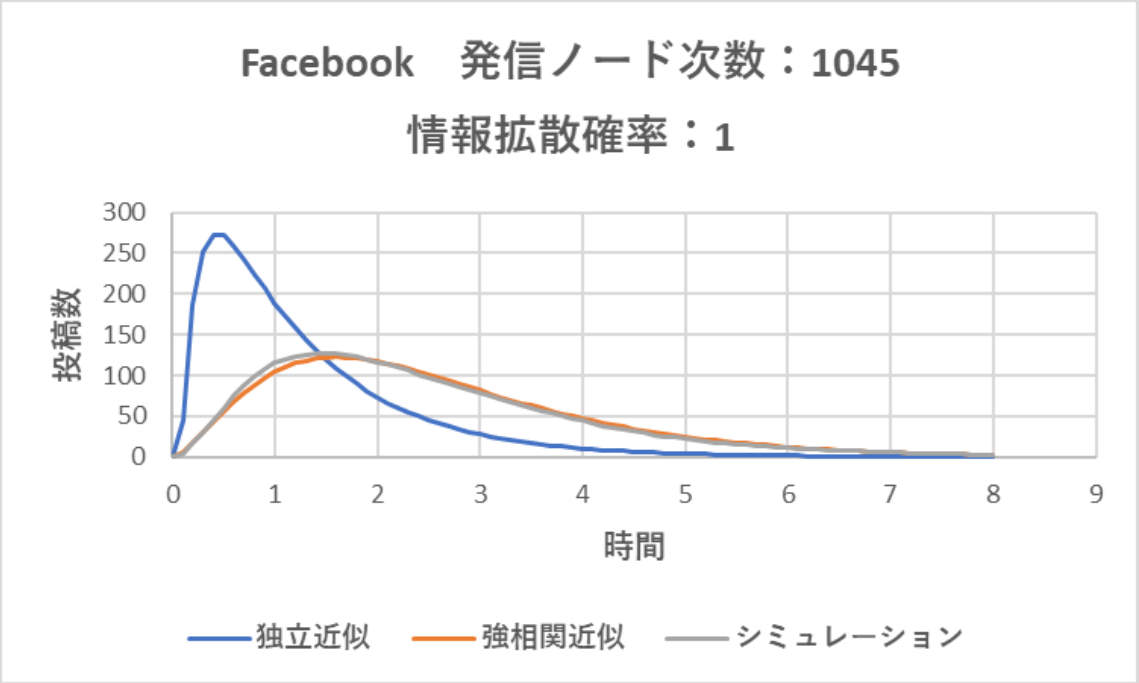

図 35　投稿数の時間推移（Facebook：発信ノード次数：1045　情報拡散確率 1）

図 30～図 35 の結果から，Facebook のトポロジーデータ上では，情報拡散確率が高くなるほど強相関近似による解析結果がシミュレーション結果を正確に再現し，独立近似による解析結果はシミュレーション結果を再現しないことが確認された．どの結果においても，強相関近似による解析結果は，シミュレーション結果を概ね正しく再現している．特に，情報拡散確率 1 (情報を受け取ったら必ず情報を拡散する)のとき，シミュレーション結果と強相関近似の解析結果はほぼ一致する．これは，Facebook の情報拡散はツリー状のネットワークのように進行していることを意味している．また，情報発信ノード次数が大きいほど，シミュレーション結果と強相関近似の解析結果は一致していくことが分かる．

4.5.2　シミュレーションと解析(Twitter)

Twitter において，発信ノード次数：41 と発信ノード次数：513 の結果をそれぞれ情報拡散確率 0.3，0.5，1.0 の場合ごとに示す．

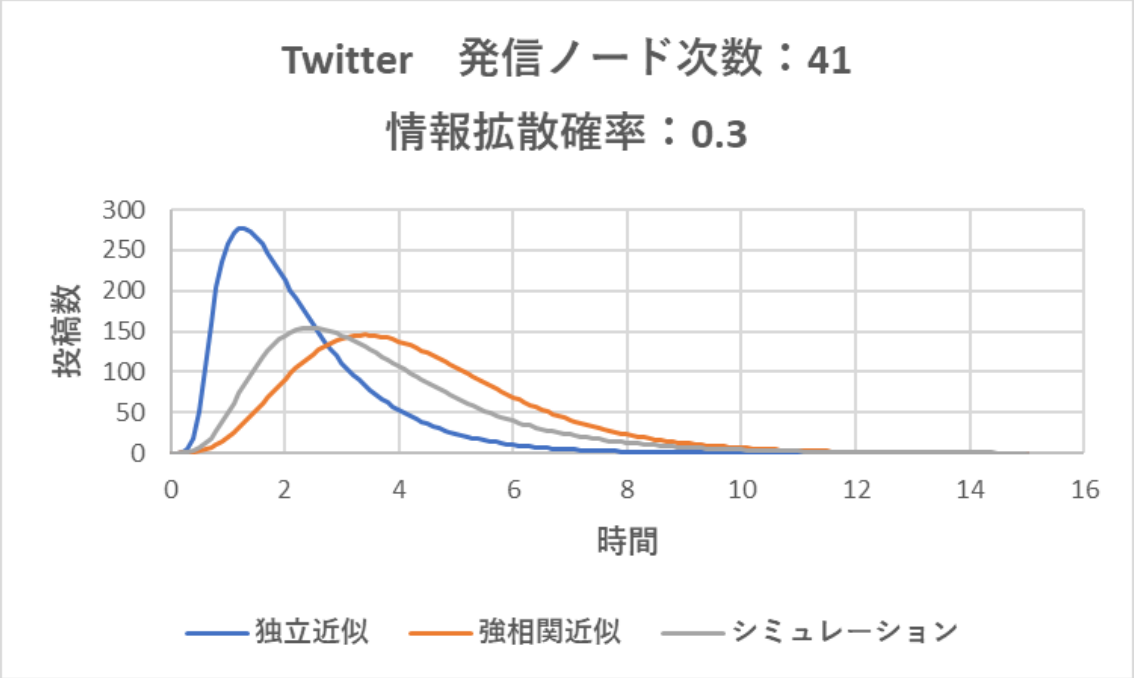

図 36　投稿数の時間推移（Twitter：発信ノード次数：41　情報拡散確率 0.3）

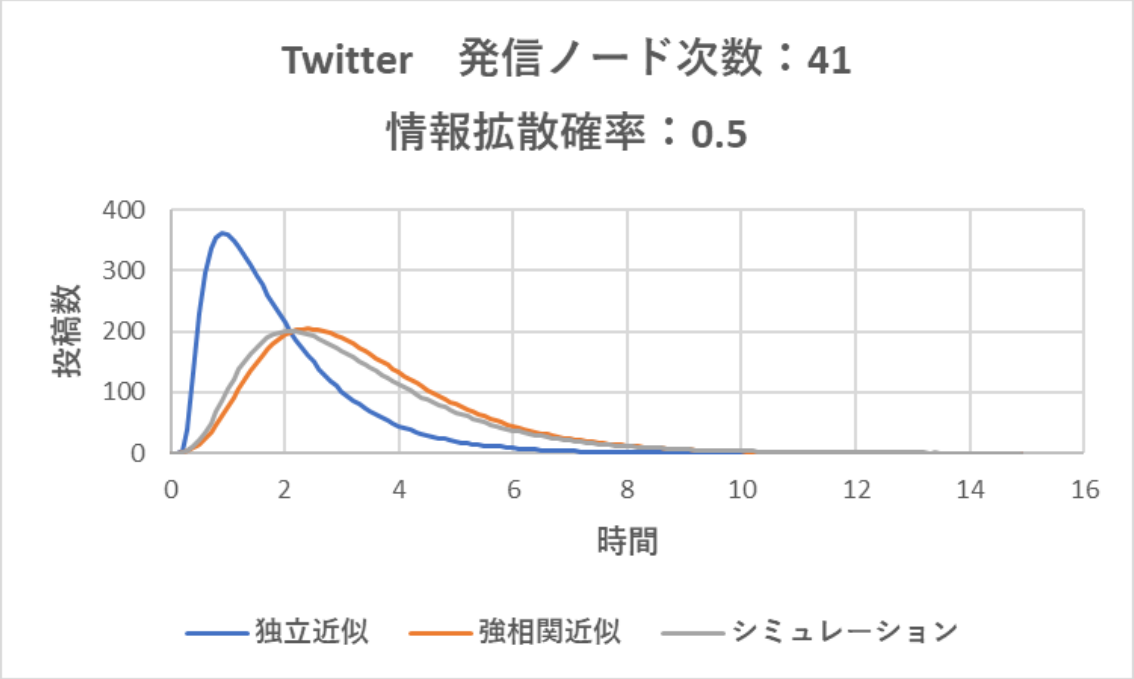

図 37　投稿数の時間推移（Twitter：発信ノード次数：41　情報拡散確率 0.5）

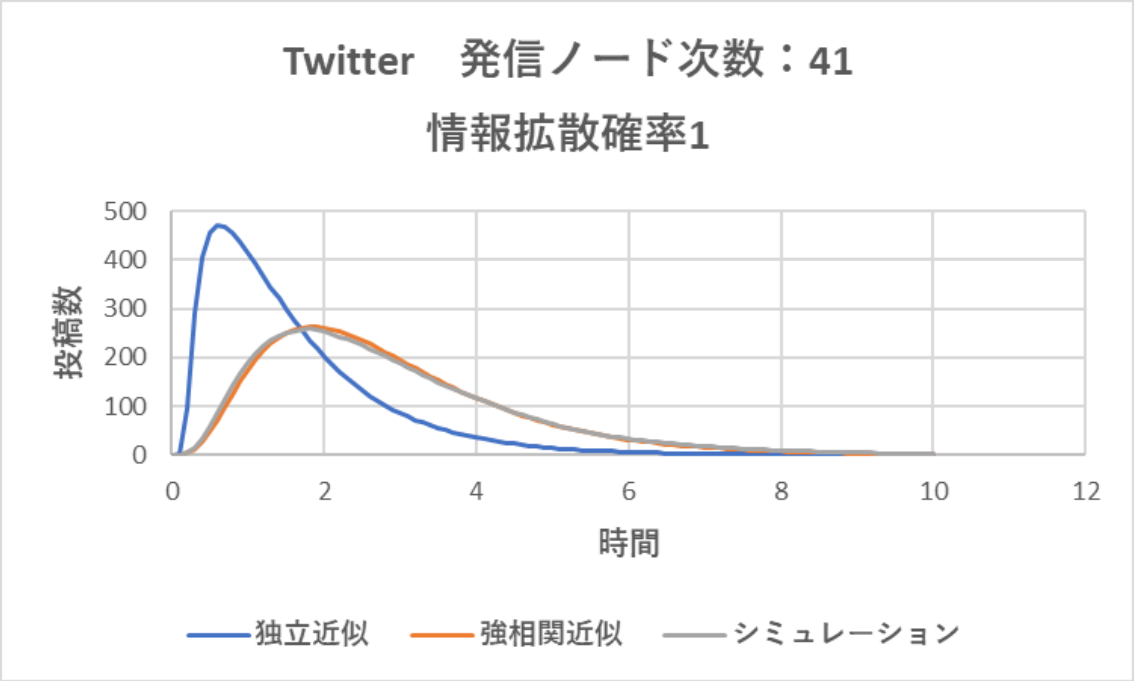

図 38　投稿数の時間推移（Twitter：発信ノード次数：41　情報拡散確率 1）

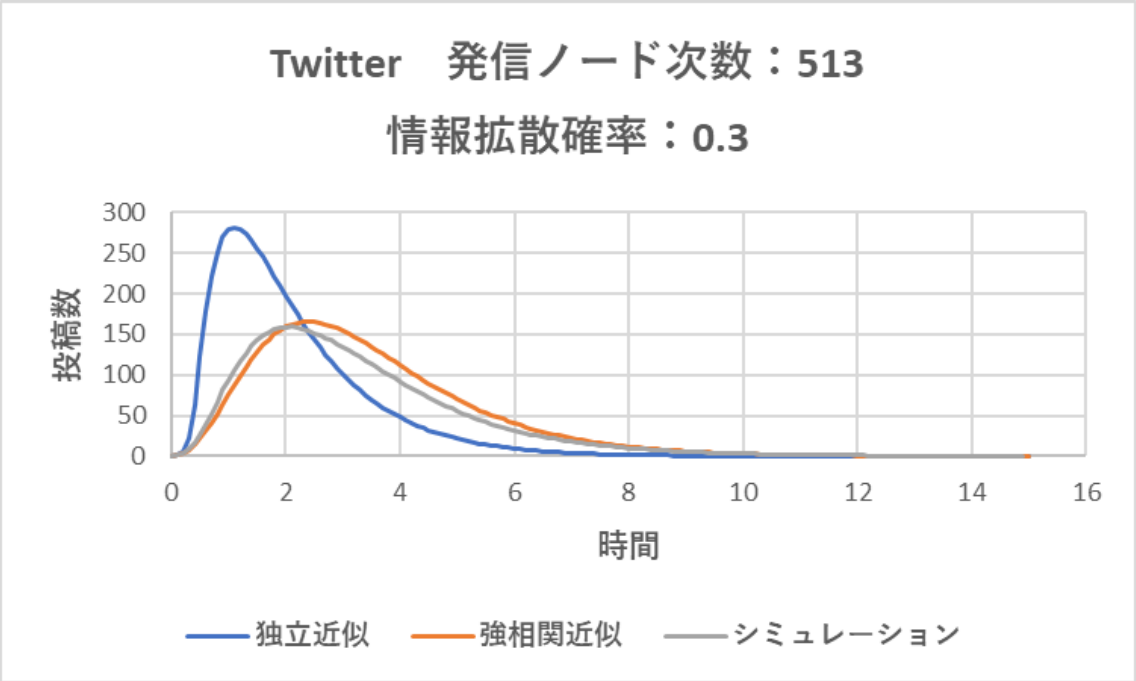

図 39　投稿数の時間推移（Twitter：発信ノード次数：513　情報拡散確率 0.3）

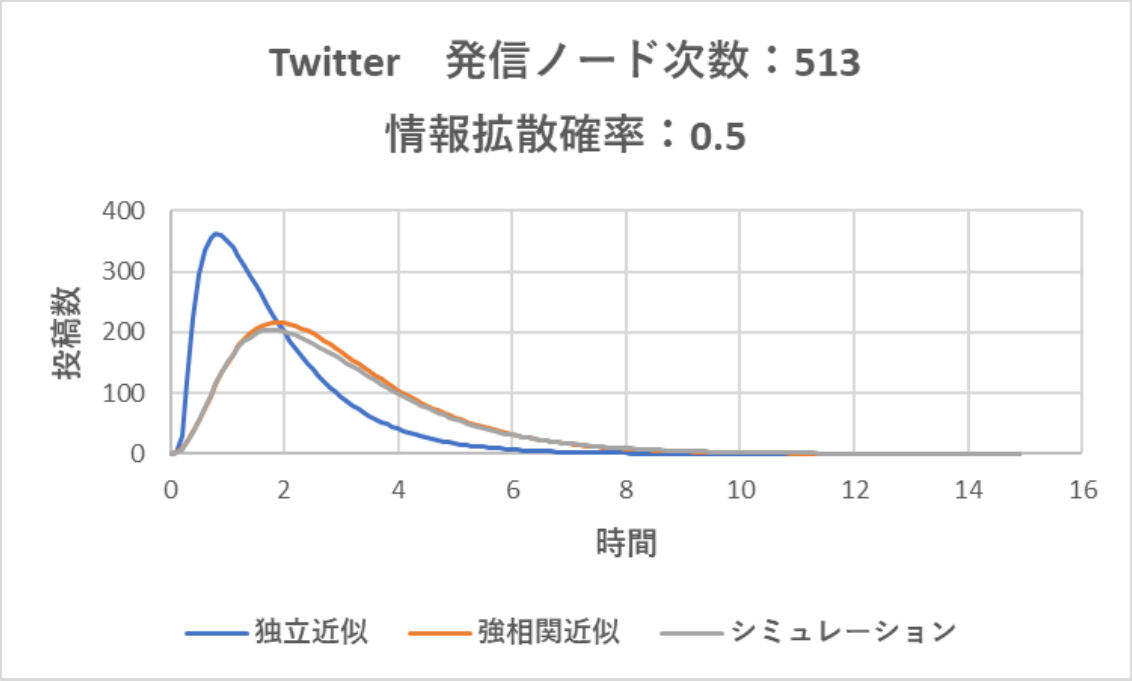

図 40　投稿数の時間推移（Twitter：発信ノード次数：513　情報拡散確率 0.5）

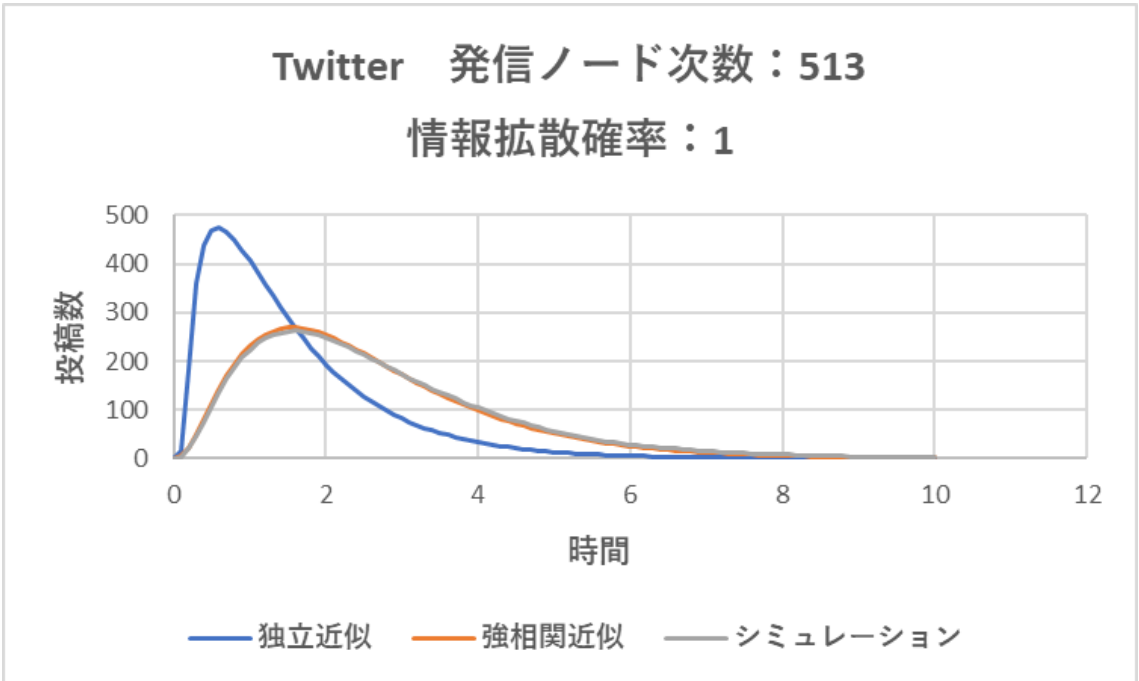

図 41　投稿数の時間推移（Twitter：発信ノード次数：513　情報拡散確率 1）

Twitter のトポロジーデータ上においても，強相関近似による解析結果は，情報拡散確率が高くなるほど正確にシミュレーションを再現する．やはり，情報発信元の次数が大きいほど投稿数のスパイクはより急峻であり，投稿数のスパイク形状の情報発信元に依存性が確認できる．また，いずれの結果においても，強相関近似により解析的に得た結果は，シミュレーション結果を正しく再現している．これは，Facebook の場合と同様，Twitter においても，情報の拡散はツリー状のネットワークの場合のように（概ね上流から下流に向かって）進行していることを意味している．

発信ノード次数が大きく，情報拡散確率が高くなるほど強相関近似による解析結果はシミュレーション結果と一致していくため，現実的には知名度の高い企業が広告を打ったとき，またいわゆる「有名人」が情報を発信したとき，強相関近似による解析は大きく効果を発揮すると考えられる．

また，解析において，微分方程式を解く際に，Y_{ij}を$\mathrm{E}[Y_{ij}]$ =0.3，0.5 で置き換えて強相関近似で評価した結果と比較した．Twitter のトポロジーデータ，発信ノード次数は 41 とした．

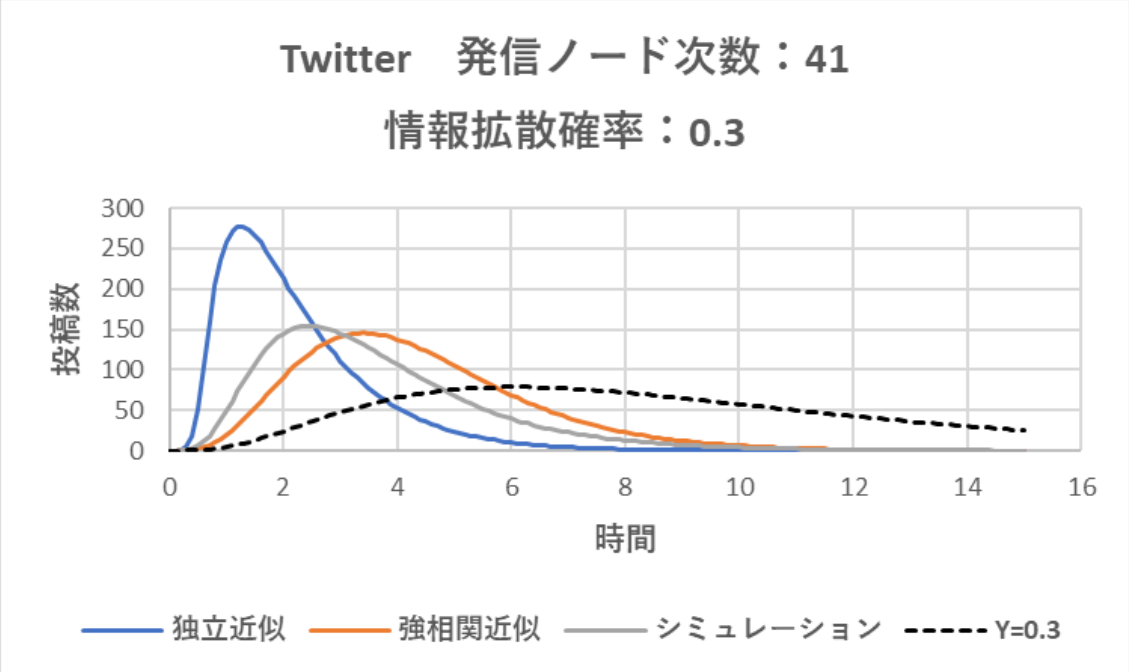

図 42　投稿数の時間推移（Twitter：発信ノード次数：41　情報拡散確率 0.3）

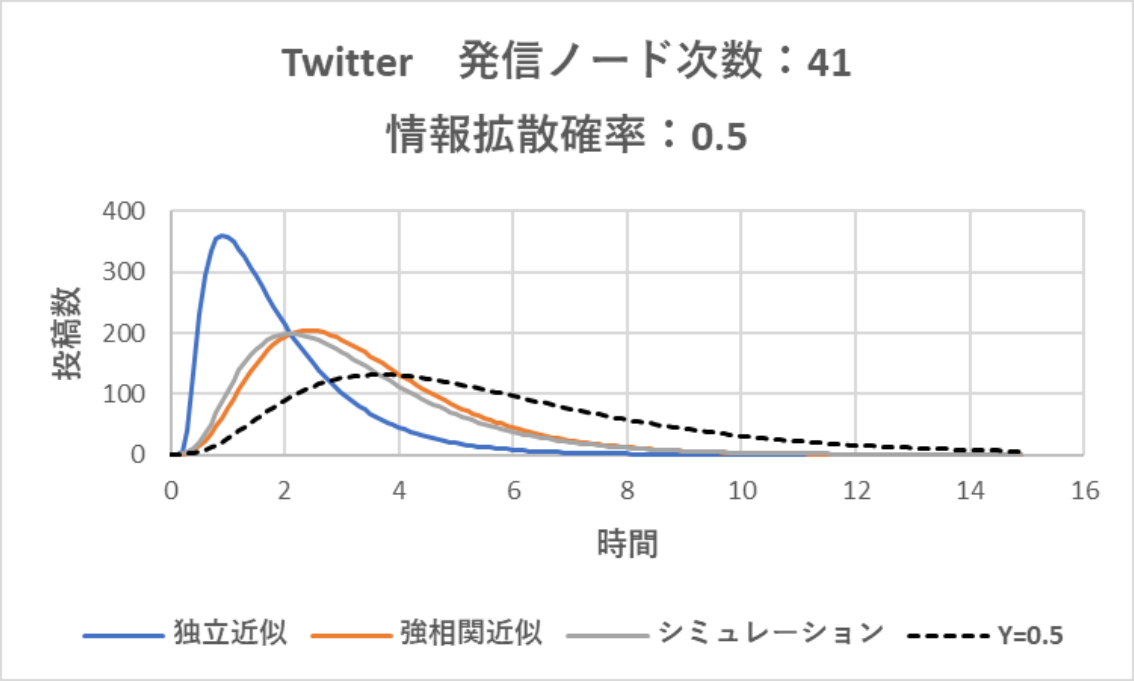

図 43　投稿数の時間推移（Twitter：発信ノード次数：41　情報拡散確率 0.5）

図 42，図 43 の結果から，Y_{ij}を$\mathrm{E}[Y_{ij}]$で置き換えて強相関近似で解析した結果は，Yの実現値を 1 つ選んで解析した結果と比較するとシミュレーション結果と大きく違う．このことから，Y の実現値を 1 つ選んで解析を行う方法は情報拡散確率が低い場合には使えないが，Y_{ij}を$\mathrm{E}[Y_{ij}]$で置き換える方法より有効であると考えられる．

第 5 章　結論

本研究では，SNS への投稿件数のスパイク現象を，ネットワーク構造を陽に取り入れた SIR モデルにより分析した．特に，モデルをシミュレーションで評価するだけでなく，SNS の投稿数の時間推移を解析的に評価する手法について考察した．その結果，SNS の投稿数のスパイク現象はモデル上で再現できること，ネットワーク構造を陽に考慮することが重要であること，強相関近似を用いることで解析的な評価が可能であること，その際に確率変数 Yを 1 実現値で代表させる手法が有効であることなどを明らかにした．

また，情報拡散確率が低い場合には，確率変数 Yの選び方(グラフ上のリンクの選び方)によって大きく解析結果が変わってしまい，シミュレーション結果とのフィッティングに利用できなかった．この「ネットワーク上で情報が拡散しなくなる最も大きい確率」＝「臨界確率」について，今後検討する必要があると考えられる．

今後は，実際のツイート数の時間推移データとのフィッティング，複数の情報が相互に影響しながら同時に拡散する場合のモデル化，SNS ネットワーク上の臨界確率等について検討を行いたい．また，投稿数のスパイクが複数個生じる現象を説明するために，拡散の途中で情報の突然変異が生じ，突然変異した情報が新規情報として新たに拡散していく現象をモデルに取り入れる予定である．

謝辞

本研究を進めるにあたり，幾度となく丁寧なご指導を頂いた塩田茂雄教授に感謝致します．また，日常の議論を通じて多くの知識や示唆を頂いた塩田研究室の皆様に感謝致します．

参考文献

[1]長尾将宏，長尾智晴，“Twitter を用いた株式市場の変動予測”，第 76 回全国大会講演論文集，2014(1)，pp．377-378，2014-03-11

[2] “Google Trends” https://trends.google.co.jp/trends/　2017/11/2 閲覧

[3] R. Pastor-Satorras, C. Castellano, P.V. Mieghem, andA. Vespignani, “Epidemic processes in complex networks, ”Reviews of Modern Physics, vol.87, pp.926–979, 2015.

[4] R. Anderson and R.M. May, Infectious Diseases in Humans, Oxford University Press, Oxford, 1992.

[5] M. Boguna and R. Pastor-Satorras, “Epidemic spreading in correlated complex networks, ” Phys. Rev., vol. E66, p.047104, 2002.

[6] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos, “Epidemic thresholds in real networks, ” ACM Transactions on Information and System Security, vol.10, no.4, pp.13.1–13.26, 2008.

[7] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, “Patterns of cascading behavior in large blog graphs, ” Proceedings of the 2007 SIAM International Conference on Data Mining, 2007.

[8] Y. Okada, K. Ikeda, *et al.*, “SIR-extended information diffusion model of false rumor and its prevention strategy for twitter, ” Journal of Advanced Computational Intelligence and Intelligent Informatics, vol.18, no.4, pp.598–607, 2014.

[9] J. Cheng, L. Adamic, J. Kleinberg, and J.Leskovec, “Do cascades recur?, ” WWW’16, 2016.

[10] D. Kempe, J. Kleinberg, and E. Tardos, “Rise and fall patterns of information diffusion: model and implications, ” KDD’03, pp.137–146, 2003.

[11] D. Watts and P. Dodds, “Influentials, networks, and public opinion formation, ” Journal of Consumer Research, vol.34, no.4, pp.441–458, 2007.

[12] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, “Outtweeting the twitterers - predicting information cascades in microblogs, ” WOSN’10, 2010.

[13] N. Barbieri, F. Bonchi, and G. Manco, “Topic-aware social influence propagation models, ” Knowledge and Information Systems, vol.37, no.3, pp.555–584, 2013.

[14] Y. Matsubara, Y. Sakurai, B. Prakash, L. Li, and C. Faloutsos, “Maximizing the spread of influence through a social network, ” KDD’12, pp.6–14, 2012.

[15] “Stanford large network dataset collection.” http://snap.stanford.edu/data/.